Satta
13.6.06
COK-H05937-20-KP2565

THE JOURNAL OF PHILOSOPHY

20

VOLUME CIII, NUMBER 1
JANUARY 2006

page

5 *Actualist Essentialism and General Possibilities* Ori Simchen
27 *Counterpart Theory, Natural Properties, and Essentialism* Todd Buras

BOOK REVIEWS

43 Harry G. Frankfurt: *The Reasons of Love* Niko Kolodny
51 MEMO TO AUTHORS
52 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Christopher Peacocke, Carol Rovane, Achille C. Varzi. *Consulting Editors*: David Albert, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2006

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

KP 2565

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.
POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2006 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University. It is typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

THE JOURNAL OF
PHILOSOPHY

VOLUME CIII, NUMBER 1
JANUARY-DECEMBER 2006

Edited by
Bernard Berofsky, Akeel Bilgrami
John Collins, Arthur C. Danto
Kent Greenawalt, Patricia Kitcher
Philip Kitcher, Isaac Levi, Wolfgang Mann
Mary Mothersill, Christopher Peacocke
Carol Rovane, Achille C. Varzi

Consulting Editors
David Albert, James T. Higginbotham
Charles D. Parsons, Wilfried Sieg

Managing Editor
John Smylie

Published by The Journal of Philosophy, Inc.

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor. Please consult our website, www.journalofphilosophy.org, for more information about our copyright policies.

© Copyright 2006 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and other mailing offices. Please send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. Fax: (212) 932-3721

THE JOURNAL OF PHILOSOPHY
VOLUME CIII, NO. 1, JANUARY 2006

ACTUALIST ESSENTIALISM AND GENERAL POSSIBILITIES*

> There is always something which underlies, out of which the thing comes to be, as plants and animals come to be out of seed. The things which simply come to be do so some of them by change of shape, like a statue, some by addition, like things which grow, some by subtraction, as a Hermes comes to be out of the stone, some by composition, like a house, some by alternation, like things which change in respect of their matter. All things which come to be like this plainly come to be out of underlying things.[1]

> For we call a thing something, when it is that thing in actuality, rather than just in possibility.[2]

> Modalities in their primary use concern counterfactuals about actual objects, and to reintroduce possibilia is to run counter to the admonition of Russell that we "retain our robust sense of reality."[3]

---

*This paper started out as comments on Russellian themes in the work of Ruth Marcus presented before the Philosophy of Language Workshop at the University of California, Los Angeles, on the occasion of the Centenary of Bertrand Russell's "On Denoting" (*Mind*, XIV (1905): 479–93), with Marcus herself as commentator. I want to thank members of the audience for their reactions at the time, especially Joseph Almog, Tony Anderson, David Kaplan, and Marcus. I also owe thanks to Andrew Irvine, Mohan Matthen, Shelly Rosenblum, Steven Savitt, and Christopher Stephens for further comments and exchanges regarding earlier versions of this paper. I owe a special debt to Almog, whose incisive criticisms led me to set some of the original ideas aside for a different occasion.

[1] Aristotle, *Physics*, Books I and II, William Charlton, trans. (New York: Oxford, 1992), 190b3–9.

[2] *Physics*, 193b7–8.

[3] Ruth Barcan Marcus, *Modalities* (New York: Oxford, 1993), p. 197.

0022-362X/06/0301/5–26

Here is a common conception of modality. Possibilities are spoken of in many ways. Suppose I say that it is possible for me not to be typing now, or, in the subjunctive, that I might not have been typing now. Such a possibility, pertaining as it does to particular things—me, the present moment—is a particular possibility. It concerns how the relevant particulars might have been had they been different from the ways they are. Other possibilities are general possibilities, such as the possibility that there be someone or other not typing at some point or other in time. In contrast to particular possibilities, general possibilities do not pertain to particulars. They obtain or not regardless of what particular things there happen to be. Then there are the mixed possibilities, such as the possibility that there be someone other than me typing in my office now.[4]

Such a broad characterization of the variety of possibilities is often thought to be beyond dispute. Indeed, it is so widespread as to be considered part of the data that any metaphysics of modality will have to account for. Nevertheless, I believe it is mistaken.

During the infancy of the subject in its contemporary guise it was assumed that possibilities are all general. A possibility such as me not typing now was thought to consist in the nonanalyticity of the contradictory, namely, me typing now. So possibility was a matter of consistency. In the language of intensions, the possibility in question reduced to the consistency of the co-instantiation of the intensions of being me, of not typing, and of being the present moment. But with the rising influence of Saul Kripke's *Naming and Necessity*[5] came an important qualification. Possibilities concerning particulars are not to be construed along such generalist lines. The possibility of me not typing now is not reducible to matters of consistency. There is me and there is the present moment. Given what we are, certain relational properties are applicable to us while others are not. These are matters determined by what we are, not by whether or not the intensions representing us are jointly consistent with the intension representing the relational property of not typing at a time. This means that possibilities concerning particulars are not cheaply had by some consistency considerations at the level of representational surrogates, where the latter are understood to be constitutionally independent of the things represented.[6] Particular possibilities are afforded to us, rather, by what the relevant things in the world already are and so are grounded in the things themselves and not in some representational surrogates.[7]

---

[4] I reserve the familiar terminology of "de dicto" and "de re" to mark a distinction in forms of expression rather than in what is being expressed. See final section for a detailed discussion of the familiar distinction.

[5] Cambridge: Harvard, 1980.

And yet general possibilities are still treated along traditional lines. They are still regarded as aloof in not depending on what things in the world there are. The possibility that there be a talking donkey, for example, is generally perceived as determined independently of what things there are, by the compatibility of being a donkey and being a talker. However such compatibility is ultimately unpacked, it is supposed to be independent of how matters stand with actual donkeys. Specifically, if it is possible that there be a talking donkey, then this is not determined as a general matter by what is possible for the donkeys, namely, by whether it is possible for any of them to talk. On many versions of the present conception, if it turns out that it is possible that there be a talking donkey, then this will certainly be evidenced by how something is, be it an actual donkey, a nonactual donkey if such items are permitted by the overall view, or an actual thing other than any actual donkey if it is impossible for any actual donkey to talk and the overall view disallows nonactual donkeys. But this is all just a trickle-down effect of the compatibility of the properties of being a donkey and being a talker. It is the compatibility that is primary in the order of explanation of the relevant general possibility.

And yet it behooves us to ask: How can any possibility regarding donkeys, however general, enjoy such independence from how things are with the donkeys? In what follows, I will make a case for the claim that no such possibility can. General possibilities are not fundamentally different from particular possibilities in being independent of what things there are. General possibilities concern portions of the world that are perhaps more scattered, more spread out in space and time, than particular possibilities, but they still concern things in the world. Whether or not there is a talking donkey is whether or not a certain fact obtains. This fact pertains to certain things—the donkeys—and it is determined by whether or not any of them talk. None do. Whether or not it is possible that there be a talking donkey is whether or not a certain counterfact obtains. It, too, pertains to the donkeys. Is it possible for any of them to talk?—Surely not. So it is impossible that there be a talking donkey.

---

[6] How to think of such matters of independence between representations and things represented is a rather subtle issue. For an extended discussion, see Ori Simchen, "On the Impossibility of Nonactual Epistemic Possibilities," this JOURNAL, CI, 10 (October 2004): 527–54.

[7] For present purposes we count David Lewis's counterparts as representational surrogates and so lump his treatment of *de re* modality together with pre-Kripkean accounts such as Rudolf Carnap's. See chapter 4 of Lewis, *On the Plurality of Worlds* (New York: Blackwell, 1986), and Carnap, *Meaning and Necessity: A Study in Semantics and Modal Logic* (Chicago: University Press, 1947). There are still other options for thinking about the determinants of particular possibilities that are neither the generalism (consistentism) of a Carnap nor the particularism (thing-dependence) outlined above. One might suppose, for example, that particular possibilities are irreducible to general ones and yet hold that what is possible is not determined by what the particulars themselves are, but, rather, by whether the intensions applying to them are mutually consistent. Thanks to Almog for discussion here. On such a view, the particulars are needed to fill in the slots, as it were, of the possibilities in question, but whether or not the possibilities obtain is a matter determined solely at the intensional level. To make this more vivid, consider for example the way in which the impossibility of a particular leaf being at one and the same time green all over and red all over is often assumed to derive from the general mutual exclusion of being green and being red.

However, there may be a different question lurking in the background of the question whether there might have been a talking donkey. Sometimes we mean to consider whether or not the world might have gone in such a way as to generate a talking donkey. This is, strictly speaking, not a question about possible being but one about possible becoming.[8] Is it possible for past conditions of the world to have eventuated in a talking donkey? Might things of the past become a talking donkey? This is certainly a different question from whether or not it is possible that there *be* a talking donkey. And the answer to this latter question is most likely in the negative as well, but not because of how things are with the donkeys, but, rather, because of how things were with early propagators of the donkeys.

In short, if (a) nothing is a possible talking donkey, and (b) nothing could have engendered a talking donkey, then there is no sense in which there might have been a talking donkey. In what follows, I broach this general outlook by exploring its ramifications for a familiar problem in contemporary metaphysics.

I. ACTUALISM VERSUS POSSIBILISM

There is a problem discussed in the modal metaphysics literature that is supposed to afflict what has come to be known as "actualism," a position that is otherwise thought to be attractive to many metaphysicians. Very briefly, *actualism* is the view according to which there are no nonactual, merely possible individuals. Its dual, *possibilism*, is the view according to which besides actual individuals there are merely possible individuals, individuals such as additional human beings to those who have appeared at some point or other in the history of the world. Admittedly, putting the contrast between the two positions in this way is of limited utility. For even this rough sketch, with its heavy reliance on the loaded notion of *individual*, is couched in difficult doctrines about being that were originally inspired by the model theory of first-order logic[9] and then transposed into the rather different context of the model theory of quantified modal logic (QML). But this much will have to be taken for granted in order to state the problem that will act as our springboard. The problem is really a problem about how to maintain the actualist insistence that there are no merely possible things in the face of properties that are both actually uninstantiated and cannot be had contingently. It is commonly supposed, for example, that essential properties cannot be had contingently. What essentialism amounts to is even more difficult and controversial than what actualism amounts to. But one thing that essentialism implies, at least initially and before excessive sophistication sets in, is that an essential property such as being a human being is not contingently instantiable. So it is necessary that if anything at all *can* have such a property, then it *must* have it.[10]

---

[8] For more on the distinction between being and becoming and their different modes, see below.

The final detail required to introduce the problem for actualism has to do with the type of interpretation of QML favored by actualists. An actualist would hold that the intended interpretation of QML is one whereby the set consisting of all and only actual individuals is assigned to the actual world as its domain of quantification, and subsets of it—perhaps proper, perhaps not—are assigned as domains of quantification to other possible worlds. This is supposed to reflect the actualist insistence that counterfactual alternatives to the way the world is do not include alternative individuals to those there are in actuality.

For present purposes, being as they are nonmathematical ones, we may circumvent the question of which semantics—a fixed-domain one or a variable-domain one—is a preferable choice for QML. It is easy to verify informally that if the intended interpretation of QML is as described above, then the Barcan Formula,

$$\text{(BF)}\ \Diamond\exists x\phi x\rightarrow\exists x\Diamond\phi x$$

---

[9] Cf. W.V. Quine, "On What There Is," *The Review of Metaphysics*, II (1948): 21–38.

[10] This is not to say that the notion of essence is itself modally reducible. For a rejection of the modal reducibility of essence see Almog, "The What and the How," this JOURNAL, LXXXVIII, 5 (May 1991): 225–44, and Kit Fine, "Essence and Modality," *Philosophical Perspectives*, VIII (1994): 1–16.

will come out true, never mind whether it is valid or not. And now here is the problem. Consider the following assumptions:

(1) 'ϕ' cannot be had contingently (so '$\Box\forall x(\Diamond\phi x \rightarrow \Box\phi x)$' is true).[11]
(2) It is possible that there be a ϕ (so '$\Diamond\exists x\phi x$' is true).
(3) Nothing ϕs (so '$\neg\exists x\phi x$' is true).

These seem to be assumptions that an actualist can accept. But from (1), (2), (3), and (BF), a contradiction follows. For by (2) and (BF), something possibly ϕs ('$\exists x\Diamond\phi x$' is true). Let *a* be such a thing ('$\Diamond\phi a$' is true). By (1), *a* necessarily ϕs ('$\Box\phi a$' is true), and so, *a* ϕs ('ϕa' is true), and so, something ϕs ('$\exists x\phi x$' is true), contradicting (3). Of course a possibilist will deny the relevant instance of (BF) and the anti-essentialist will deny the relevant instance of (1). The problem is a problem for actualist essentialism.

Now consider any ordinary claim that there might have been a ϕ. According to the general outlook on general possibilities espoused above we understand such a claim in one of two ways: either it is possible that there be a ϕ, in which case it is possible for some thing to ϕ; or else it is possible that something become a ϕ, in which case it is possible for something (or some things) to have given rise to a ϕ. But (2) in the above argument is already a disambiguation of the ordinary claim that there might have been a ϕ—it is the first sense that is at issue. So now, if there really is no ϕ, and if nothing could ϕ as a matter

---

[11] This characterization of ϕ not being contingently instantiable might be contested on the general grounds that concreteness is contingent, and so, to the extent that ϕ pertains to concreta, (1) will have to take the contingency of concreteness into account. For a general defense of the contingency of concreteness, see Bernard Linsky and Edward Zalta, "In Defense of the Contingently Nonconcrete," *Philosophical Studies*, LXXXIV (1996): 283–94. On such a view, to say that a property ϕ pertaining to concreta is not contingently instantiable might be understood as the modified claim that it is necessary that if anything possibly ϕs, then necessarily it ϕs if it is concrete. Let '*C*' be the concreteness predicate. The suggested replacement of (1) will be:

$$(1') \ \Box\forall x(\Diamond\phi x \rightarrow \Box(Cx \rightarrow \phi x)).$$

(The embedded consequent is just the Linsky/Zalta analysis of *x* bearing such a ϕ essentially.) While replacing (1) with (1′) certainly avoids the contradiction discussed in the text, (1′) in the case of concreteness itself becomes the valid

$$\Box\forall x(\Diamond Cx \rightarrow \Box(Cx \rightarrow Cx)).$$

This strongly suggests that concreteness cannot be thought of as on a par with other properties pertaining to concreta. But it is difficult to see why we should think of concreteness in this special way. (To claim, along with the Linsky/Zalta line, that such a treatment of concreteness is recommended by an interpretation of the "simplest" QML surely puts the formal semantic cart before the metaphysical horse.)

of mere possibility without already ϕ-ing, then it turns out that it is impossible that there be a ϕ after all. In other words, (2) is false. A proper appreciation of what it means to deny that there are merely possible things requires us to deny (2) in cases where the property in question is uninstantiated and nothing could serve as a possible instance without already being an instance. In this way, actualist essentialism properly construed is left untouched by the above argument.

This is an underexplored way out of the problem.[12] It turns out to be metaphysically loaded in a particular way that is likely to offend the aesthetic sense of those with a taste for subtle alternations in the desert landscape of formal frameworks. But it is, I believe, well motivated quite independently of the particular way the present problem is posed. We fill in the details by examining cases.

II. UNICORNS

Let us begin with the highly compressed and notoriously difficult discussion of unicorns in Kripke's *Naming and Necessity*.[13] In that work Kripke famously proclaims that it is not the case that there might have been unicorns. As might be expected, it is highly controversial how to understand Kripke's general grounds for this denial of possibility. There are at least two competing readings of his scattered pronouncements on the subject. One ("definitionalist") faction understands the denial of the possibility as due to some lack in a "real definition" that would specify the species, even as a matter for possible instantiation.[14] Perhaps a real definition in the case of a putative species would be a formula specifying a putative DNA sequence. And surely such a specification is missing for the case at hand. Another ("worldly") faction understands the denial of the possibility that there be unicorns as due to an absence in real-world grounds for the generation of such a thing as a species, even as a matter for possible instantiation.[15] On such a rendering, a species is itself something that is generated by certain worldly conditions at a particular moment in world history. *De dicto* claims involving species terms are understood as *de re* claims about the relevant species. To say, for example, that some horses are striped is to say something about *Equus caballus* itself, to wit, that some of its specimen are striped. Ditto with respect to saying that it is possible for horses to be striped, or possible that there be horses.[16] And clearly a recasting of the claim that there might have been unicorns along such lines is unavailable due to the absence of the species in question.

---

[12] A prominent actualist strategy for accommodating properties that are both uninstantiated and cannot be had contingently is the reductionist effort to make possibilist discourse actualistically acceptable. See, for example, Fine's "Postscript," in Arthur Prior and Fine, *Worlds, Times and Selves* (Amherst: Massachusetts UP, 1977), pp. 116–61. Such a program includes a reduction of *de re* modal claims involving the relevant properties so that no mere possibilia will be called upon for their possible instantiation. Much effort and ingenuity have gone into such reductionist programs but from the present perspective they concede too much to possibilism at the very outset. Possibilist intuitions need not be indulged. They demand, rather, to be traced all the way back to their origins and confronted at the appropriate level. This is the approach to be attempted here.

[13] See *Naming and Necessity*, pp. 23–24 and 156–58.

[14] For more on the notion of real definition and its assimilation to the notion of essence, see Fine's "Essence and Modality." Fine, however, opts for an actualist reduction of possibilist discourse; see note 12.

Between the two readings of Kripke on unicorns the second seems closer to the overall spirit of *Naming and Necessity* while the first seems closer to the letter of the relevant passages, but this is, again, a matter of considerable dispute. Be that as it may, closer scrutiny of Kripke's own treatment of the unicorn case reveals an interesting lacuna that can be filled by the present understanding of general possibilities. Kripke claims that it is not the case that there might have been unicorns. This is issued as an antidote to the Carnapian tendency to suppose that while there certainly are no unicorns, nor have there ever been any, it is nevertheless possible as a purely general matter that there be unicorns. (Carnap would say that the property of being a unicorn, while *F*-empty, is not *L*-empty.[17]) Kripke's strategy is as follows. He first advances the claim that *if* it is possible that there be the species of unicorns, *then* something, some particular species or other ("actual or possible"), is such that it is possible for that thing to be the species in question. Next he asks: Which particular thing is it for which it is possible to be the species of unicorns? No such thing is to be found. It follows that it is not possible that there be the species in question to begin with.

---

[15] See Almog, "The Structure-in-Things: Existence, Essence and Logic," *Proceedings of the Aristotelian Society*, CIII (2003): 197–225.

[16] Perhaps on such a view '$\exists x \phi x$', with 'ϕ' a species term, may be parsed as: "There is an x such that x is of the kind ϕ," with "the kind ϕ" understood to refer to the relevant species. Cases where we engage in kind-talk without being in a position to refer to the relevant kinds are exceptional, but the example of ununseptium discussed below may be a case in point. In such a case the above treatment can apply to the constituents of a descriptive reduction of the term. Let "ununseptium" receive the reduction "an atomic element whose instances have a nucleus containing 117 protons." Then x is of the kind ununseptium just in case for all y such that y in an atomic constituent of x, there is a z and a w such that z is a nucleus of y, w is a proton, and there are 117 ws in z. Let us set aside the delicate issue of how to parse the numerical attribution "There are 117 ws in z." The natural kind and relation predicates "y is an atomic constituent of x," "z is a nucleus of y," and "w is a proton," are to be understood as "y bears the relation *being an atomic constituent of* to x," "z bears the relation *being a nucleus of* to y," and "w is of the kind *proton*," while the constituent expressions "the relation *being an atomic constituent of*," "the relation *being a nucleus of*," and "the kind *proton*" are understood to refer to the relevant relations and kind, respectively.

[17] See, for example, *Meaning and Necessity*, pp. 64–68.

A peculiar feature of this argument is that it treats the target claim that there might have been unicorns as equivalent to the claim that it is possible that there be the species of unicorns. Yet the claims are surely not equivalent. In the language of properties, the first involves the property of being a unicorn while the second involves the property of being identical with the species of unicorns. To see that they are not equivalent, consider a skeptic about species in general who accepts Kripke's conditional claim that *if* it is possible that there be the species of unicorns, *then* some species or other is possibly the species of unicorns. Let the skeptic think that the consequent is false, and therefore the antecedent, because there are no such things as species, and thus nothing for which being identical with the species of unicorns is even a possibility. Such a skeptic can hold that it is possible that there be unicorns despite thinking that it is not possible that there be the species of unicorns.[18]

Of course Kripke's own motivations for switching from a consideration of the general possibility that there be unicorns to the particular possibility that there be the species of unicorns is anyone's guess. But the following hypothesis seems plausible. Kripke's original metaphysical insight that possibilities involving membership in natural kinds depend on the actual world to supply the kinds was adulterated by an approach to general possibilities that makes them independent of what things there are. He then sought to assimilate general possibilities involving kinds to particular possibilities in order to preserve the metaphysical insight, for regarding the latter he could claim that they surely depend on the actual world to supply the kinds just as much as the possibility of me not typing now depends on the actual world to supply the relevant particulars, me and the present moment.

But with our general approach to general possibilities we need not force general possibilities involving kinds into the mold of particular possibilities. Switching back from the property of being (identical with) the species of unicorns to the property of being a unicorn, we may paraphrase Kripke's overall argument as follows: *if* it is possible that there be unicorns, *then* something is a possible unicorn. (This shows the germ of truth in (BF) as applied to the present case.) And yet no particular thing is a possible unicorn. So it is not the case that it is possible that there be unicorns to begin with.

---

[18] On such a view the existence of organisms does not entail the existence of species of organisms.

In keeping with our general approach to general possibilities we concur in denying the general possibility that there be unicorns. If such a thing were possible, then something would be a possible unicorn. But nothing is a possible unicorn. For if "unicorn" is a putative species term, then to be a unicorn cannot be had contingently. This means that if anything were a possible unicorn, then it would already be a unicorn. Yet there are no unicorns. We conclude that it is not possible that there be unicorns to begin with.

Let us go back to the problem for actualist essentialism discussed in the previous section. As it stands, the present strategy for the actualist accommodation of properties that are both uninstantiated and cannot be instantiated contingently faces a potential challenge in the form of the following question: Are there compelling grounds for thinking that all cases involving uninstantiated-yet-noncontingently-instantiable properties are ones for which (2) in the argument of the previous section is false? The example of unicorns in this context can seem distracting. Unicorns are commonly considered paradigm cases of the mythic or the fictive and it may seem that due to this fictive character alone the property of being a unicorn is impossibly instantiated. Whether or not this is a proper view of the case is an interesting question that cannot be pursued here, but in any case we should strive to enrich our diet of philosophical examples. To solidify our sense that the falsity of (2) generalizes to other cases of uninstantiated-yet-noncontingently-instantiable properties, we move on to consider two types of cases that might seem more challenging: one involving an actual past kind and one involving a potentially future kind. Doing so will pave the way for a more nuanced understanding of the present approach to actualist essentialism. The rest of the paper will be devoted to an exploration of the implications of the falsity of (2). It will turn out that properly motivating such a strategy requires alterations in currently entrenched ways of thinking about the subject that are more radical than initial appearances may have revealed.

III. ACTUAL PAST KINDS

Consider the dodo bird. This species, *Didus ineptus*, native to the island of Mauritius, is now extinct. In keeping with our general approach we ask: Is it really impossible that there be dodos? There once were dodos. There are no more. In order to sustain the present existence of dodos, history would have needed to take a different course than the one actually taken. Are there compelling reasons specific to the case for denying the possibility that there be dodos? Because, once again, if there are no such reasons, then from the essentiality of being a dodo, the possibility that there be dodos, and the actual absence of dodos, a contradiction follows by the general conception of general possibilities espoused here. Also, the two readings of Kripke's grounds for the claim that it is impossible that there be unicorns are not obviously applicable here. There is certainly no lack in a real definition to specify the species *Didus ineptus*. Alternatively, there is no absence in the worldly grounds for generating the species, for dodos did once roam the island. Had history only taken a somewhat different course, the species would not have become extinct.

At this point an obvious retort suggests itself. The alleged problem posed by dodos depends on a presentist reading of the argument for the contradiction. The challenge is to say how it is that it is impossible that there be dodos at present, given that being a dodo is essential to whatever has it and given that there are no dodos at present. But such a temporal reading of the argument leading to the contradiction is independently unwarranted. The dodo challenge dissipates once we understand that argument in a temporally unrestricted way. (2) is indeed true. It is possible that there be dodos because there were once dodos. But (3) is false and for the same reason. There once were dodos, so it is not the case that there are no dodos in a temporally unrestricted way. In short, while history might have taken a different turn, first on the island of Mauritius and then elsewhere as well, this is, strictly speaking, irrelevant to the possibility that there be dodos. Actual past dodos stand witness to the general possibility that there be dodos understood in the temporally unrestricted way. Accordingly, there is also something that is a possible dodo. Any actual dodo of the past will do.

While this dissolution of the dodo challenge is effective as far as it goes, there are signs of trouble lurking in the wings. Consider the temporally indexed property of being a presently living dodo. It may seem *prima facie* possible that there be a presently living dodo. But if that is so, then by our approach something in the world must stand witness to such a possibility. Let us grant that to be a presently living dodo is not essential to whatever has it. For example, a mature dodo in 1650 might have failed to survive into 1650, which is a reliable indicator for it not being of its nature to be a dodo *in 1650*. So (1) is false and the problem for actualist essentialism is avoided. And yet a question persists regarding the general possibility that there be a presently living dodo. What is it by virtue of which it is possible that there be a presently living dodo?— Presumably, something for which it is possible to be a presently living dodo. But it is difficult to see how anything could be a possible presently living dodo without also being a possible dodo. If so, then by the lights of the above reply to the dodo challenge the thing which is a possible presently living dodo is going to be an actual past dodo. But then again when we originally pondered the general possibility that there be a presently living dodo we meant that history might have taken a different course from the one actually taken so that *Didus ineptus* would not have become extinct. We did not mean this to involve the possible survival of any specimen of the species into the present day, say possibly frozen and then thawed and revived. This latter, somewhat exotic scenario represents a highly remote sense in which there might have been a presently living dodo. What about the more mundane sense, the sense that does not require the possible survival of any past dodo? Is it possible that there be a presently living dodo that is not a past dodo? It certainly can appear so. But if so, then by the lights of our general treatment of general possibilities something would have to be a possible presently living dodo that is not a past dodo. And what might such a thing be?

IV. POTENTIALLY FUTURE KINDS

A second type of case that seems to present a challenge to the general idea that (2) is false when the property in question is both uninstantiated and cannot be had contingently are potentially future kinds. Consider ununseptium, atomic number 117. We suppose that this as-of-yet undetected element is nowhere to be found in actuality. We run the argument again. Suppose it is possible that there be ununseptium. Then by the present conception of general possibilities something or other must stand witness to such a possibility by being a possible instance. Let *u* be such a thing. Well, being an instance of ununseptium cannot be had contingently. So if *u* is a possible instance of ununseptium, then it is an instance of ununseptium as a matter of necessity. And so *u* is an instance of ununseptium in actuality, contradicting the initial assumption that there is no ununseptium.

Let us digress for a moment and set aside the question whether it is generally possible that there be ununseptium. Some may feel torn by their wish to endorse the claim that something is a possible instance of ununspetium while affirming the essentiality of being an instance of ununseptium. Such a tendency merits a digression, for it holds an important clue to filling in the details of the overall solution to the problem offered here.

Scientists are interested in producing an instance of element 117 and there seems to be no good reason for thinking that they are incapable of success in their efforts. Let us assume that this is really so, that it is indeed possible that scientists produce an instance of ununseptium. What would happen in the event of success is the production of an instance of ununseptium. More specifically, a certain transformation of some instance of another atomic element—let us say ununhexium, atomic number 116—into an instance of ununseptium would take place. A new atomic element would then come into being.

Now consider the instance of ununhexium for which it is possible to get transformed into an instance of ununseptium.[19] There may be some inclination to say that it is a possible instance of ununseptium. But if it is a possible instance of ununseptium, then we have a straightforward counterexample to the purported essentiality of being an instance of ununseptium. For essentiality here implies that being an instance of ununseptium is not a contingent matter. Thus, if it is possible for something to be an instance of ununseptium, then it is already actually so. But by hypothesis there are no instances of ununseptium.

One option here might seem to be a revision of our understanding of the implications of essentialism. For example, regarding an essential property ϕ we might want to replace the demanding equivalences

possibly-ϕ-ing $\leftrightarrow$ necessarily-ϕ-ing $\leftrightarrow$ actually-ϕ-ing

with the truncated

necessarily-ϕ-ing $\leftrightarrow$ actually-ϕ-ing.[20]

But a moment's reflection will reveal that this will not help. For suppose that an instance of ununhexium is transformable as a matter of mere possibility into an instance of ununseptium. And suppose that we opt for the suggested replacement. Still, how can the instance of ununhexium be even possibly an instance of ununseptium? Consider: it is actually an instance of ununhexium. Being an instance of an atomic element other than ununseptium is essential for it, and therefore necessary for it. But it is impossible for anything to be an instance of more than one atomic element. So the instance in question is not a possible instance of ununseptium after all.[21]

---

[19] We might prefer to regard the sum of a given instance of ununhexium and a given proton as that for which it is possible to get transformed into an instance of ununseptium. For present purposes this does not matter.

[20] In the language of QML '$\Box\forall x(\Diamond\phi x\rightarrow\Box\phi x)$' would be dropped in favor of '$\Box\forall x(\phi x\rightarrow\Box\phi x)$'.

[21] A general point looms here. Properties that invariably pertain to what things are, such as being an instance of an atomic element or being of a certain species, cannot be had as a matter of mere possibility. Here is how we can ground this further. First, we suppose that it makes good sense to speak of properties excluding one another. So, for example, being green all over excludes being red all over. And so, being necessarily green all over excludes being possibly red all over. Next, we suppose that for any property ϕ that invariably pertains to what things are, any non-ϕ has properties pertaining to what it is (that is, there is no "prime matter") among which will be found a property ψ that excludes ϕ in the above sense. Now let *o* be a non-ϕ that is nevertheless possibly ϕ. Well, *o* is already ψ for some property ψ that pertains to what *o* is and excludes ϕ. So in order for *o* to possibly ϕ it would have to be possible for *o* to be both ϕ and ψ, which is impossible.

The correct alternative is to deny flat-out that the instance of ununhexium that would get transformed into an instance of ununseptium can survive such a transformation.[22] From this point of view it is simply wrong to speak of whatever might get transformed into an instance of ununseptium as a possible instance of ununseptium. The pre-transformation instance I^{PRE} is an instance of ununhexium. Being an instance of ununhexium pertains to what it is and is therefore necessary for it. This means that it is impossible for it to be an instance of ununseptium. Once the transformation sets in, I^{PRE} peters out of existence and the distinct I^{POST} emerges, for which being an instance of ununseptium is essential and therefore necessary.

Going back to the main problem then: Is (2) in the case of ununseptium true? Is it possible that there be an instance of ununseptium? For all that has been said so far, an actualist essentialist who happens to be a definitionalist about essence might be forced to say that it is. There certainly seems to be no lack in a real definition to specify ununseptium, so in this respect the present case is unlike that of the general possibility that there be unicorns. Some special pleading by an actualist essentialist of such a persuasion will be needed in order to establish that it is not possible that there be ununseptium after all. This brings us to our preferred approach to the denial of (2).

V. THE LARGER CONTEXT OF DENYING (2)

On the present conception, to be an actualist is not a matter of simply opting for this or that interpretation of QML, for this or that set of modal principles. Nor is it a matter of accommodating full-fledged possibilist discourse.[23] The actualist should not feel particularly pressed to provide actualistically acceptable paraphrases for each and every claim put forward by the possibilist. Much of what passes for modal platitudes in recent discussions of the subject is already thoroughly tainted by possibilism. To be an actualist requires dealing with the metaphysics directly and letting the logic track the metaphysics rather than the other way around.

---

[22] As in the following passage from Marcus: "Being gold or being a man is not accidental.... No metaphysical mysteries. Such essences are dispositional properties of a very special kind: if an object had such a property and ceased to have it, it would have ceased to exist or it would have changed into something else. If by bombardment a sample of gold was transmuted into lead, its structure would have been so altered and the causal connections between its transient properties that previously obtained would so have changed, that we would not reidentify it as the same thing"—"Essential Attribution," this JOURNAL, LXVIII, 7 (April 8, 1971): 187–202, on p. 202.

[23] See note 12.

By way of introduction to the present approach we subscribe to the following outlook. We begin with kinds. Natural kinds are themselves actual things that are in a certain sense on a par with their instances. As with any actual thing, they tend to bear essential relations to aspects of their emergence in actual world history. The relation of a natural kind to its instances is not a relation of a universal to a particular falling under it, where the former is understood to be constitutionally independent of its applying to anything. Nor is it a relation between an impure set and its members. A species, for example, exists through the existence of, and interactions among, its specimen, not unlike the way a higher organism exists through the existence of, and interactions among, its cells. Something similar may be said about a chemical compound. Being a kind of configuration of atomic elements, it exists through the existence of, and interactions among, its elements. And the same holds for the atomic elements themselves, those being kinds of configurations of elementary particles. The atomic elements exist through the existence of, and interactions among, the elementary particles. And so it goes.[24]

We can see the relevance of this general outlook to the problem posed by actual past species by considering a related case. Suppose it is possible for my paternal grandfather, who passed away over a decade ago, to have survived to this day. If 'ϕ' stands for the property of being a living red blood cell originating from his own marrow, then given the possibility of my grandfather's survival we wish to say that there might have been instances of ϕ. How so? We assume that the property in question cannot be had contingently. One sense of the claim, as seen above in the case of dodos, is that some past instances of ϕ might have survived, in which case there is a straightforward sense in which it is possible that there be ϕs: past actual instances of ϕ are also possible instances of ϕ. But that is not what we would ordinarily mean by the claim that there might have been ϕs. What we would mean is that there might have been instances of ϕ due to there being some actual conditions underlying my grandfather's past existence that might have conspired, together with some counterfactual conditions underlying his survival to this day, to generate novel instances of ϕ. So it is not that it is possible for some non-ϕ-ing thing to instantiate ϕ in the requisite sense without being a past actual instance. It is, rather, that under the assumption of the possibility of my grandfather's survival it is possible that some past things generate novel instances of ϕ. And so, some past things are such that they might have generated instances of ϕ. It is in this modified sense, then, that we need to understand the claim that there might have been ϕs. It is possible that some plurality of things of the past, under suitable counterfactual conditions, give rise to novel instances of ϕ by way of generating them.

---

[24] There appears to be a difference between kinds sustained by interactions among their instances and kinds sustained by interactions among constituents of their instances. The topic of kinds is complex and very much in need of further study which cannot be undertaken here.

Recall our query regarding the possibility that there be a presently living dodo that is not a past dodo. If things had gone differently on Mauritius in the seventeenth century, *Didus ineptus* would not have become extinct. So we feel impelled to affirm the general possibility that there be a presently living dodo. And yet nothing is a possible presently living dodo, for if something were such, then it would also be a possible dodo. And given the essentiality of being a dodo, such a thing would already be a dodo. An unsatisfying rejoinder, as we saw above, is to say that it is some past actual dodo that is possibly a dodo. This is unsatisfying, once again, because the survival of specimen from the seventeenth century is not what we mean when we say that it is possible for the species not to have become extinct. And in any case the possibility of survival of past specimen is on far shakier grounds than the possibility of the survival of the species itself.

A better response is needed, and it is this: let us consider the time before the species became extinct at which possible history of dodo survival splinters off from actual history of dodo extinction. Let us suppose that the actual conditions underlying dodo existence at that time and eventuating in dodo extinction are $c_1,\ldots,c_i,c_{i+1},\ldots,c_n$, and let us suppose that some portion of these, $c_1,\ldots,c_i$, together with some counterfactual conditions $c_{i+1}^*,\ldots,c_n^*$, would have eventuated in the continued survival of the species to this day. This is the relevant sense in which there might have been dodos now. Some actual conditions underlying past dodo existence, together with some counterfactual conditions underlying dodo survival, might have given rise to presently living dodos. In other words, it is possible that some plurality of things of the past, under the right conditions, would have generated presently living dodos. The claim that it is possible that there be presently living dodos that are not past dodos is rejected in favor of the claim that it is possible that there be generators of novel presently living specimen of the species.

Turning now from a *de dicto* mode of expression to a *de re* mode, is there anything of which we can say that it is a possible presently living dodo? No, there is not. But if we reject the *de dicto* claim that it is possible that there be presently living dodos that are not past dodos in favor of the claim that it is possible that something of the past generate presently living dodos, then we also get a *de re* dual. Some aspects of $c_1,\ldots,c_i$—aspects of the actual history of the species *Didus ineptus*—might have given rise to contemporary instances of the species. What might those aspects of $c_1,\ldots,c_i$ be?—Things of the actual past, which, under suitable counterfactual conditions, would have generated dodos and sustained dodo existence to this day. In short, things of the actual past, perhaps the last generation of dodo eggs that was actually destroyed, had an unfulfilled *potentiality* to become dodos, which would have generated later dodos in turn, and so on and on all the way up to the present time. This does not require that anything be a possible presently living dodo that is not a past dodo. In short, potentiality without the corresponding possibility is the key to a properly motivated denial of (2) of the problem for actualist essentialism.

Before we continue, a general reminder about being and becoming—and their modes: possibility and potentiality—is in order. Consider a seed of an oak tree, or a fertilized human egg. Of such things we say that they are potentially an oak tree and potentially a human being, respectively. By this we mean that the seed might become something else, an oak tree, and that the fertilized egg might become something else, a human being. By this we do not mean that the seed might be an oak tree, or that the fertilized egg might be a human being. That would be to say that something that is not a tree is possibly a tree. But being a tree is essential to whatever has it, so if the seed is possibly an oak, then it is already an actual oak. And yet the seed is most certainly not an oak. An oak seed is no possible oak and a fertilized human egg is no possible human. But an oak seed and a fertilized human egg are *potentially* an oak and a human, respectively. Potentiality is a matter pertaining to what the seed and the egg might become. Potentiality is possible becoming.

So again, why is it that we are inclined to affirm that there might have been presently living dodos? Not because we think it possible for anything to *be* a presently living dodo. This option either violates the essentiality of being a dodo or else represents a possibility far too remote to be taken seriously. We think that there might have been a presently living dodo because we think it possible that some plurality of things of the past *become* a presently living dodo. And so we think that something (some things) is (are) such that it is possible for it (them) to have become a presently living dodo. But (2) is false: it is not the case that it is possible that there be a presently living dodo. The falsity of (2) removes the threat of the contradiction.

KP 2565

Similar remarks apply to the possibility that there be ununseptium. We think that there might have been ununseptium even in its actual absence, not because we think it possible that there *be* an instance of ununseptium. It is impossible that there be an instance of ununseptium because it is impossible for anything to be an instance of ununseptium. And yet, something might have become an instance of ununseptium. Perhaps some instance of ununhexium might have become an instance of ununseptium if conditions were right. This is no violation of the essentiality of being an instance of an atomic element. In the event that we succeed in actualizing ununseptium, something—an instance of ununhexium, say—will go out of existence and something else will come into existence. And if it is possible that something become an instance of ununseptium, then something is such that it is possible for it to have become an instance of ununseptium. Once again, due to the falsity of (2) no contradiction is forthcoming here either.

As a further illustration of the present approach we return to the example familiar from the literature on actualism of the possibility of having a brother despite being actually brotherless. Let us assume that given that I have no brother, nothing is such that it could have been a brother of mine.[25] The present suggestion is that if it seems that I might have had a brother, then this is only because it seems possible that things of the past become a brother of mine, which is to say that things of the past are such that it is possible for them to have become my brother. This is not to say that it is possible that anything be a brother of mine. The latter is indeed impossible. But we can certainly suppose it possible that genetic materials from my parents generate another male offspring. If so, then we can also say of such genetic materials that they might have become a male sibling of mine without making the preposterous claim that the genetic materials themselves somehow constitute a possible brother of mine. In short, our general diagnosis as to why we are drawn to the relevant instance of (2) in cases where it is impossible for anything to ϕ is that we think it possible that things might have gone in such a way as to generate a ϕ. It is possible that something *become* a brother of mine. And this certainly does not entail that it is possible that anything *be* a brother mine. The quantifier ranges over everything. And there is just no possibility that being my brother be instantiated by anything.

---

[25] The noncontingency in this case is best construed as that of the relational property *being a brother of*. If any two things are possibly related in this way, then they are related in this way as a matter of necessity and so already related in this way.

VI. *DE DICTO* AND *DE RE*

To see why such a treatment is appropriate for the cases at hand we need to step back and consider any uncontroversial claim schematizable by '$\Diamond\exists x\phi x$', say that possibly something other than the thing occupying the corner of this room might have occupied it instead. What grounds do we have for thinking any such *de dicto* claim is true? Intuitively, we think that something other than the actual occupier is a possible occupier of the corner. In other words, we think that of all the things there are, something else might have occupied the corner. And we think this because we think that occupying the corner is not precluded by what something other than the actual occupier is.

An alternative and less intuitively appealing explanation for why think that the *de dicto* claim is true is an explanation along the following lines. Let the actual occupier be *o* and let its spatial extension and orientation relative to the corner of the room be *E*. We might think that the properties of having-oriented-shape-*E* and being-other-than-*o*—the latter understood to involve some purely qualitative specification of what it is to be *o*—are possibly co-instantiated. And why should we think this?—Because such co-instantiation gives rise to no contradictions within the overall story of what is projected to be the case under relevant generalizations about the subject matter at hand.

So we have two types of explanation, a particularist one and a generalist one. The former type of explanation is generally preferable and is the more natural of the two, but a full comparison between them would extend beyond the scope of this paper.[26] For present purposes the following remarks should suffice. The first take on *de dicto* modal claims considers their truth or falsity to be determined by what things there are, while the second approach considers them not to be so constrained. This difference is reflected in a contrast between two ways of conceiving of counterfactual possibility more generally. On the one approach, counterfactual possibility concerns what things would have been like had they been different than the ways they are. Some of them would have been different from the ways they in fact are. Others would not have come to be in the first place. But in general, counterfactual evaluation concerns the world *qua* things. Not so on the second approach. The second approach considers counterfactual possibility to be a matter of counterfactual co-instantiation of properties. And this is supposed to be independent of what things there happen to be. Strictly speaking, on the second approach counterfactual possibility is not determined by what things would have been like had they been different from the ways they are, but, rather, by which properties would have been co-instantiated. From this perspective *de re* modal claims are exceptional and thus deserving of special treatment. Typically, they have been thought to reduce to *de dicto* locutions.

---

[26] For further discussion of the issue as it arises for *de re* modal claims see my "On the Impossibility of Nonactual Epistemic Possibilities," pp. 534–40.

The second approach has certainly enjoyed, and continues to enjoy, widespread appeal, and it is worth pausing to reflect on how this came about. It may be traced back to two sources. One source is a certain way in which the medieval distinction between *de dicto* and *de re* has been appropriated in recent times. According to this received view, *de dicto* modality is modality pertaining to *dicta*, where the *dicta* themselves are understood to be fundamentally thing-neutral.[27] In the wake of the so-called New Theory of Reference we ought to remain highly suspicious of such thing-neutrality.[28] A different and more proximate source of the appeal of the second approach to *de dicto* modal claims is the contemporary philosophical influence of the variable-domain semantics for QML introduced by Kripke. Taken as a guide to the metaphysics of modality, Kripke's formalism turns what things there are, what things fall in the range of the quantifiers, to be a matter that is itself open to counterfactual variation. An important outcome of this general license has been the prevalent philosophical sentiment that letting what things there are constrain what we claim to be possible *de dicto* is to succumb to a parochial bias. But we need to consider afresh whether such an attitude is justified. After all, within Kripke's formal apparatus counterfactual scenarios are counterfactual by stipulation. Possible worlds (the set Kripke labels '*K*') are just indices of which the actual world ('*G*') is one. And the latter is distinguished from the rest by nothing more than being so designated. Specifically, it is not distinguished from the rest by what constitutes it. Yet surely counterfactual evaluation properly so-called is a consideration of what the world would have been like had the facts been different in specified ways. And while variation in what facts there are may require *some* variation in what things there are, it is a mistake to think of the former variability as entirely unconstrained by what things there are. If it were so, then we would be hard pressed to see how modal claims concern the world at all. For we think that what things there are in the world is of primary importance to any consideration of what the world might have been like for the purposes of counterfactual evaluation. In what, after all, does the identity of the world consist if not in what things constitute it? The license to leave things out of consideration in counterfactual evaluation is a license to leave the world itself out of such consideration. To put the matter somewhat dramatically, it is but a short step from taking Kripke's formalism with metaphysical seriousness in letting *de dicto* modal claims be unconstrained by what things there are to regarding the world as a point of evaluation, a mere index.

---

[27] Apparently this is not how the staunch nominalist Abelard thought of the distinction between *de re* and *de dicto* he introduced. According to William Kneale, Abelard did not regard modality *de dicto* as genuine modality at all unless it derived from modality *de re*. See Kneale, "Modality *De Dicto* and *De Re*," in Ernest Nagel, Patrick Suppes, and Alfred Tarski, eds., *Logic, Methodology and the Philosophy of Science: Proceedings of the 1960 International Congress* (Stanford: University Press, 1962), pp. 622–33, especially at p. 624.

[28] See my "On the Impossibility of Nonactual Epistemic Possibilities" for more details. It should be noted that while some adherents of the second approach to *de dicto* modal claims (for example, Kneale) do appear to suppose not only that counterfactual co-instantiation of properties is independent of what things there are but also that the properties themselves are constitutionally independent of what things there are, the second commitment is not compulsory. The point in the text is meant as a diagnosis of a tendency to think of *de dicto* modality in a particular way.

Of course more needs to be said against the second approach to *de dicto* modal claims and in favor of the first. This will have to be deferred to another occasion. The foregoing should at least give some indication of how the argument would run. In the meantime we opt for the first approach. With this type of understanding on hand we can easily see why it is impossible that anything be a brother of mine. Such a thing is impossible because nothing in the world is such that it might have been my brother. No*thing* has a nature compatible with being a brother of mine, so the claim that it is possible that something be a brother of mine is false. But while the original claim is false, another claim in the vicinity is true, namely, that it is possible that something become a brother of mine. And the latter claim is true by virtue of the past existence of some genetic materials from my parents (or what have you).

Could we run the argument against actualist essentialism with becoming my brother instead of with being my brother? Let 'ϕ' be being my brother and 'ψ' be becoming ϕ. As we saw above, it is impossible that anything ϕ. But it is possible that something (a plurality of things) become ϕ, and so possible that something (some things) jointly ψ. So something (some things) possibly ψs (jointly ψ). And yet nothing ψs. An analogous problem would arise for this case only if ψ-ing could not be had contingently. But it had better turn out not to be so. (Which is not to say that *possibly* ψ-ing, that is, *potentially* ϕ-ing, that is, potentially being my brother, may not essential to whatever has it.) For suppose that ψ-ing could not be had contingently. Well, ψ does not actually apply to anything, so presumably it would be impossible for it to apply to anything. Generalizing from this case would yield that there are no unactualized potentialities. And yet unactualized potentialities abound.

Finally, it may be thought that the real problem cases for actualist essentialism are ones in which what seems intuitively possible is for some property cleverly constructed so as not to be capable of being instantiated by anything actual to be instantiated. Such a construction need not employ the term "actual" itself. For example, it may be thought possible that there be an additional carbon atom to those there are, that is, additional to those appearing at some point or other in the history of the world. What makes such cases interesting is their high degree of generality. The properties are defined so that their possible instantiation would be as free as can be from what particular things there are. But from the present perspective this type of neutrality is an illusion. If the possibility envisioned is general enough, then it may seem difficult to pinpoint any particular item (or items) that might become an instantiator of the property. For example, if the possibility that there be an additional carbon atom is indeed general enough, then it may seem that there cannot be any item(s) to point to as the thing(s) that might become an additional carbon atom. The present proposal is that what might have generated an additional carbon atom is something occupying a relevant, sufficiently comprehensive portion of space-time. But again, it is impossible that anything *be* an additional carbon atom.

ORI SIMCHEN

University of British Columbia

COUNTERPART THEORY, NATURAL PROPERTIES, AND ESSENTIALISM*

David Lewis advised essentialists to judge counterpart theory a false friend.[1] He also argued that counterpart theory needs natural properties.[2] I argue that natural properties are all essentialists need to find a true friend in counterpart theory.

If I am right about this, counterpart theorists cannot have natural properties and keep their anti-essentialism, too. They will have to choose between the benefits of adding natural properties to their ontology and the costs of essentialism. Essentialists, on the other hand, need not choose between essentialism and counterpart theory. They may explore a counterpart-theoretic articulation of essentialist doctrines. Perhaps essentialists will not find the counterpart-theoretic version adequate; perhaps counterpart theorists will not find it objectionable. My aim in this paper is not to evaluate the counterpart-theoretic version of essentialism from either the counterpart theorist's or the essentialist's perspective. My aim is rather to establish that the addition of natural properties to counterpart theory is sufficient to produce a counterpart-theoretic version of essentialism, whatever its merits may be.

Counterpart theory plus natural properties equals essentialism. This is my thesis. My argument has two premises:

> *First Premise*: If there are natural properties in the counterpart theorist's ontology, then a determinate relation of comparative overall natural similarity obtains between any number of objects independently of the way they are conceived, described, or referred to.
>
> *Second Premise*: If a determinate relation of comparative overall natural similarity obtains between any number of objects independently of the way they are conceived, described, or referred to, then things have real essences, that is, counterpart theory is essentialist.

---

* For comments which contributed significantly to the development of this paper, I am grateful to Michael Della Rocca, Laurie Paul, Tad Brennan, Keith DeRose, and Troy Cross. I am also grateful to the philosophy department colloquium at Baylor University, and especially Zach Manis and David Alexander, for helpful discussions of this material.

[1] Lewis, "Postscripts to 'Counterpart Theory and Quantified Modal Logic'," in his *Philosophical Papers*, Volume 1 (New York: Oxford, 1983), pp. 39–46, here p. 42.

[2] Lewis, "New Work for a Theory of Universals," *The Australasian Journal of Philosophy*, LXI (1983): 343–77; reprinted in his *Papers in Metaphysics and Epistemology* (New York: Cambridge, 1999), pp. 8–55.

0022-362X/06/0301/27–42

The argument these premises offer for my thesis is valid, and the truth of the second premise is, I think, granted by Lewis. My main task is thus to defend the truth of the first premise. To set the stage, I begin by explaining why I think Lewis grants the second premise, as well as his reasons for adding natural properties to his ontology in the first place.

I

Lewis's counterpart-theoretic ontology consists of every possible individual together with every class composed of individuals according to the most unconstrained principles of composition.[3] Any class thus composed is identified with a property. Given the unconstrained principles governing the composition of classes, properties are extremely abundant: "Any two things share infinitely many properties, and fail to share infinitely many others."[4] Properties are also extremely undiscriminating: "Any class of things, be it ever so gerrymandered and miscellaneous and indescribable in thought and language, and be it ever so superfluous in characterizing the world, is nevertheless a property."[5]

Natural properties are a sparse and discriminating group of all the properties in the counterpart theorist's ontology. They are an "elite minority" of properties whose "boundaries are established by objective sameness and difference in nature."[6] Non-natural properties "carve reality at joints"—and every where else as well. Natural properties stick to the joints. They are "discriminatory classifications not of our own making."[7]

For my purposes, it will not matter exactly what counterpart theorists add to their ontology to make the elite minority of properties stand out from the "countless throng." For the sake of simplicity, I suppose that they posit a primitive (two-placed) resemblance relation uniting the members of the classes with which natural properties are identified.[8] I reserve the term *resemblance* (and cognates) for the unanalyzed relation that makes for naturalness. Following Lewis, I also treat the naturalness of a property as a matter of degree—

---

[3] "New Work for a Theory of Universals," p. 9.

[4] "New Work for a Theory of Universals," p. 13.

[5] "New Work for a Theory of Universals," p. 12.

[6] Lewis, "Putnam's Paradox," *The Australasian Journal of Philosophy*, LXII (1984): 221–36; reprinted in his *Papers in Metaphysics and Epistemology*, pp. 56–77, see p. 65.

[7] "Putnam's Paradox," p. 67.

[8] The other things Lewis considers positing in order to secure natural properties are: primitive naturalness of classes and genuine universals; see "New Work for a Theory of Universals," pp. 13–15.

perfectly natural properties being, roughly, classes whose members all resemble one another, less natural properties being, roughly, classes whose members do not all resemble one another.[9]

Lewis's reason for adding natural properties to his ontology is that, for the purposes of characterizing many features of the world, the vast majority of the properties things share is irrelevant. Yet there is nothing in the counterpart theorist's ontology to account for the privileged role played by a small minority of properties. To have an adequate theory of properties, the counterpart theorist must add something to the ontology that accounts for the inegalitarianism a complete characterization of the world demands. Lewis develops this line of reasoning in far more detail than we need consider here, arguing specifically that counterpart theory needs natural properties to analyze intrinsicness (and related notions), causal powers, and the content of language and thought. Of course, even fully developed, this sort of argument shows only that the philosophical price of denying natural properties is exceedingly high—too high, at any rate, for Lewis. It does not show that a counterpart-theoretic ontology is inconsistent or untenable without natural properties.

If counterpart theory without natural properties is hard pressed to account for some features of reality, it is famously resourceful in characterizing others. Indeed Lewis's primary reason for adopting such a generous domain of quantification is its resourcefulness, especially when it comes to accounting for the modal features of reality. He identifies possible worlds with nonoverlapping classes of possibilia, and treats modal operators as quantifiers over worlds. Individual x is possibly F iff there is a possible world according to which x exists and is F; and x is necessarily G iff there is no possible world according to which x exists and is not G. Since worlds do not overlap (that is, the classes of possibilia with which worlds are identified have no common members), no individual exists in more than one world. What it is for an individual from one world to exist according to another is for an individual from one world to have a counterpart in another. Thus x is possibly F iff there is a possible world in which a counterpart of x is F; and x is necessarily G iff there is no world in which a counterpart of x is not G.

So far forth there is nothing in counterpart theory inimical to essentialism. Like Lewis, I take essentialism to be the doctrine that things have real essences, that independently of the way things are conceived, described, or referred to there is a determinate fact of the matter about an object's modal properties (about the truth value of *de re* modal propositions). The essentialist doctrine is easily articulated in the counterpart-theoretic vocabulary. Any property a thing shares with all its counterparts is an essential property; the whole of a thing's essence consists of all the properties it shares with all its counterparts. Counterpart theory looses all affinity for essentialism only after Lewis specifies what it is for objects to be counterparts (and thus what it is for a property to be part of an object's essence): x is a counterpart of y iff x and y inhabit different worlds and x is overall more similar to y than anything else in x's world.

---

[9] Lewis, "New Work for a Theory of Universals," p. 13; "Putnam's Paradox," p. 65.

The truth makers for *de re* modal propositions are thus, in the final analysis, comparative overall similarity relations among inhabitants of different worlds. But for counterpart theorists, comparative overall similarity relations are extremely vague, because the similarity relation is extremely vague; and the similarity relation is extremely vague because properties are extremely abundant. Counterpart theory identifies the similarity relation, S, like any other property, with a class. To keep things simple, suppose similarity is a two placed relation, and thus identified with a class of ordered pairs, which we designate Σ. Then, for any objects x and y, x is similar to y iff the ordered pair of x and y are members of Σ. The pair $\langle x, y\rangle$ is a member of Σ iff x and y both instantiate some property, P; and x and y both instantiate some property, P, iff x and y are both members of the class with which P is identified, call it Φ. So, the ordered pair x and y are members of the class of similar things iff both are members of a class that defines a property. Given Lewis's constraints on the composition of classes, x and y are co-members of infinitely many classes. Since Lewis identifies each of these classes with a property, x and y share infinitely many properties and fail to share as many again. Consequently the extension of Σ is vague in the extreme: for any property, P, by virtue of which x and y make it into Σ, there is also a property, P', by virtue of which they do not; since for any class, Φ, of which x and y are both members, there is a Φ' that includes x but not y.

Given the counterpart theorist's account of properties, everything is similar to everything else in infinitely many respects, and there is nothing in the ontology to privilege some ways of being similar over others. Consequently, counterpart theorists can give no straight answer to questions of the form: Is x possibly F? The problem with answering simply "yes" or "no" is not that there are no mind-independent facts which constitute x's being possibly F. There are: the facts about the properties occupants of different worlds share, and the similarity and counterpart relations they bear as a result. Nor is the problem that the mind-independent facts which constitute x's being possibly *F* are, in themselves, indeterminate. The classes with which the relevant properties are identified are all perfectly determinate. The problem is that modal properties (and hence essences) are vague for the counterpart theorist.

The modal facts suffer an embarrassment of riches. The facts about similarity across worlds make virtually every *de re* modal proposition both true and false—true relative to some specifications of the way things must be similar in order to count as counterparts, false relative to others. Whether or not *x* is possibly *F* depends on which respects of similarity to *x* we take into consideration. Independently of some disambiguation of the counterpart relation, there is no one determinate answer.[10] The disambiguation required to answer such questions is, according to Lewis, indicated by context, specifically by the way we refer to *x* in thought or speech.[11] Thus things have nominal but not real essences: things do not have determinate modal properties (and *de re* modal propositions do not have determinate truth values) independently of the way they are conceived, described, or referred to.

Because of the vagueness of the similarity relation, then, counterpart theory is a contextualist version of anti-essentialism. Other forms of anti-essentialism deny that there are any mind independent facts that constitute truth makers for *de re* modal propositions. Such anti-essentialism looks to way we conceive, describe, or refer to objects—mind-dependent facts—as the only viable truth makers. Lewis's counterpart theory accepts mind-independent facts which constitute *x*'s being possibly *F*. It accepts so many viable truth makers that they alone fail to render *de re* modal propositions determinately true or false. This form of anti-essentialism looks to the mind-dependent facts—the facts about the way we conceive, describe, or refer to *x*—to select some among the many facts which are all equally legitimate truth makers. It leaves it to context to select the facts which constitute an *x*'s being possibly *F* rather than the facts which constitute *x*'s being not possibly *F* as the relevant facts for assigning truth values to claims about *x*'s modal properties, and therefore *x*'s essence.

---

[10] Lewis makes this point forcefully in several places. See *On the Plurality of Worlds* (New York: Blackwell, 1986), pp. 248–63, especially p. 251; "Postscript to 'Counterpart Theory and Quantified Modal Logic'," pp. 42–45; and "Counterparts of Persons and Their Bodies," this JOURNAL, LXVIII, 7 (April 8, 1971): 203–11; reprinted in his *Philosophical Papers*, Volume 1, pp. 47–54.

[11] For an idea of how context may function in this way, see Lewis, "Scorekeeping in a Language Game," *Journal of Philosophical Logic*, VIII (1979): 339–59; reprinted in Lewis, *Philosophical Papers*, Volume 1, pp. 233–49.

Lewis summarizes the reason he thinks counterpart theory anti-essentialist thus:

> I am by no means offering a wholehearted defense of "Aristotelian essentialism." For the essences of things are settled only to the extent that the counterpart relation is, and the counterpart relation is not very settled at all. Like any relation of comparative overall similarity, it is subject to a great deal of indeterminacy....
>
> Further, as with vagueness generally, the vagueness of the counterpart relation—and hence of essence and *de re* modality generally—may be subject to pragmatic pressures, and differently resolved in different contexts. The upshot is that it is hard to say anything false about essences. For any halfway reasonable statement will tend to create a context that (partially) resolves the vagueness of the counterpart relation in such a way as to make that statement true in that context.[12]

For my purposes, the important point of this explanation of why counterpart theory is anti-essentialist is that it relies on a tight conceptual connection between the vagueness of similarity relations and anti-essentialism. According to Lewis, things have real essences iff they have determinate specification-independent counterpart relations; and things have determinate specification-independent counterpart relations iff they have determinate, specification-independent, overall similarity relations. Because there are no such similarity relations independent of context, there are no real essences, only context-dependent (or nominal) essences.

The conceptual connections Lewis relies on to show that counterpart theory is anti-essentialist also support the truth of my second premise. Lewis and I agree about what the world must be like if counterpart theory is to be essentialist: the ontology must assign a privileged role to some determinate overall similarity relations, independently of our conceiving, describing, or referring to objects. Both before and after Lewis added natural properties to his ontology, he maintained that there are no such similarity relations in his ontology. I grant that he was right about the status of such similarities before the addition of natural properties, but not after. Thus, as advertised, my dispute with Lewis comes down to the truth of my first premise.

II

My first premise claims that natural properties entail a determinate and specification-independent relation of overall natural similarity between any number of objects. The starting point of my defense of this premise is the observation that discriminations among properties entail discriminations among similarities. Similarity, after all, is defined in terms of property sharing (which is defined in terms of class co-membership). Distinguish the properties in terms of which similarities are defined and you distinguish the similarities too. Not just any property is a natural property. So not just any similarity is similarity with respect to a natural property; or, not just any similarity is a natural similarity, where

---

[12] "Postscript to 'Counterpart Theory and Quantified Modal Logic'," p. 42.

> x is *naturally similar* to y $=_{\mathrm{df}}$ x shares one of y's natural properties ($=_{\mathrm{df}}$ x is a co-member of a natural class y belongs to).

I have defined the natural similarity relation in terms of the natural properties Lewis grants. The natural properties Lewis grants therefore are sufficient conditions of the natural similarities I have defined. My definition is not circular because the resemblance relation that identifies the natural properties invoked in the *definiens* is primitive. The definition holds of objects independently of the way they are conceived, described, or referred to, because it is not up to us, but to the distribution of the primitive resemblance relation, to determine which properties are natural; nor is it up to us, but to the principles governing the composition of classes, to determine what classes objects belong to. Also the definition is not trivially satisfied by any pair of objects, because natural properties are sparse.

But the natural similarity relation defined above is none too determinate (independently of the way the relata are conceived, described, or referred to). For the natural similarity relation is as vague as natural properties are abundant; and natural properties, though sparse compared to the "countless throng" of all properties, are still quite abundant. Natural properties are abundant because they come in many degrees. Suppose, as Lewis believes, the most natural properties are those discovered by physics.[13] This would suggest that the class of, say, charged things, would be among the most natural classes, the classes most united with respect to primitive resemblance. But co-members of the class of charged things are members of many different (and only slightly less natural) classes which are defined in terms of charge, for example, classes of sub-atomic particles; and co-members of the class of sub-atomic particles are members of many different (and incrementally less natural) classes which are defined in terms of sub-atomic particles, for example, classes of atoms. So

---

[13] "Putnam's Paradox," p. 65.

Lewis believes it goes for simplest molecules and the most complex chemical, biological, psychological, and social properties: less natural properties are related to the most natural by chains of definition, longer and longer chains of definition yielding less and less natural properties.

The result is that for any natural property two things share, there is sure to be another natural property, differing only slightly in naturalness, which the same two things do not share. For example, two things may be naturally similar with respect to being carbon atoms, and yet not naturally similar with respect to being constituents of living cells. Thus it remains indeterminate whether any pair of objects is naturally similar or not. Relative to some specifications of the natural similarity relation they probably are. Relative to others, they certainly are not. As long as it is indeterminate whether any pair of objects is naturally similar, it also indeterminate which pair of any number of objects is overall naturally similar. Thus the fact that natural properties entail natural similarities does not yet show that natural properties entail a determinate relation of comparative overall natural similarity.

But the distinction between natural and unnatural properties is not the only distinction that results from adding primitive resemblance relations to the counterpart theorist's ontology. Primitive resemblance relations not only account for the distinction between natural and unnatural properties, they also account for distinctions among degrees of naturalness. Though Lewis says little about the factors that determine just how natural or unnatural a property is, he does show how perfectly natural properties can be defined in terms of primitive resemblance.[14] What Lewis calls perfectly natural properties are perfectly natural in the sense that they are completely natural, not unnatural in anyway. A perfectly natural property, in Lewis's sense, meets the following two conditions: every thing inside the class primitively resembles everything else inside the class, and nothing outside the class is such that it primitively resembles everything inside the class and everything inside the class primitively resembles it. Perfectly natural properties contain all and only the things that bear some primitive resemblance to one another. There can be no doubt that there are such natural classes in the counterpart-theoretic ontology. This is a straightforward consequence of introducing primitive resemblance relations into the ontology, together with the unconstrained composition of classes.

---

[14] "New Work for a Theory of Universals," p. 15, n. 9.

Given the distinction between perfectly and less-than-perfectly natural properties, a determinate relation of overall natural similarity may be defined as follows. First we define perfectly natural similarity as follows:

> x is *perfectly naturally similar* to y $=_{\mathrm{df}}$ x shares one of y's perfectly natural properties ($=_{\mathrm{df}}$ x is a co-member of a perfectly natural class y belongs to).

Like the natural-similarity relation, perfectly natural similarity is ultimately defined in terms of the primitive resemblance relations Lewis grants, and it obtains independently of the way x and y are conceived, described, or referred to. But unlike simple natural similarity, perfectly natural similarity is determinate. What wrecked the determinacy of simple natural similarity was that for any natural property two things share there is another, only slightly less natural property they do not share. That is to say, the ontology left it up to us (and specifically to the context of our modal utterances) to disambiguate the borders of natural similarity. The borders of perfectly natural similarity are not ambiguous. Perfectly natural similarity has what all other similarity relations have thus far lacked: a determinate bottom-end set by features of the counterpart theorist's ontology. Two things may share countless properties that are only slightly less than perfectly natural. This certainly counts for something, but it does not count for perfectly natural similarity.

Of course, y may have no perfectly natural properties, in which case nothing will be perfectly naturally similar to y. This is not to say that nothing will be *very* naturally similar to y, or even as naturally similar to y as anything can be. Any x that shares many of y's less-than-perfectly natural properties will be very naturally similar to y; and any thing that shares all of y's less-than-perfectly natural properties will be as naturally similar to y as anything can be. But perfectly natural similarity is not (as the name may suggest) a matter of being as naturally similar as can be. Rather it is a matter of sharing perfectly natural properties. Things that do not have such properties cannot share them, and thus cannot achieve perfectly natural similarity relations, however impressive their other similarities relations.

Also, y may have more than one perfectly natural property. But we need not require x to share *all* of y's perfectly natural properties in order for x to be perfectly naturally similar to y. To do so would set the bar for perfectly natural similarity needlessly high. There may be more than one way for an x to achieve perfectly natural similarity to y, and we may sometimes want to consider only the x's that achieve perfectly natural similarity to y in one way rather than another. But this does not mean that the extension of the perfectly natural similarity relation is vague, that is, that any pair of *x* and *y* is in or out of the extension depending on how we are considering *y*. An *x* that is not a co-member of a perfectly natural class with *y* is not perfectly naturally similar to *y*, conceive or describe things as you will.

The next step is to define a relation of *comparative perfectly natural* similarity, one that specifies which pair of any number of objects is overall perfectly naturally similar.

> *x* is *overall perfectly naturally similar* to $y =_{df} x$ is perfectly naturally similar to *y*, and there is no thing, *z*, distinct from *x* and *y*, such that *z* is perfectly naturally similar to *y*, and *z* shares more perfectly natural properties with *y* than *x*.

Because overall perfectly natural similarity is defined in terms of perfectly natural property sharing, and perfectly natural property sharing is defined in terms of the primitive resemblance relations, the primitive resemblance relations Lewis introduces into counterpart theory are, once again, sufficient conditions for the relation I have defined. But the crucial point is that overall perfectly natural similarities are not, like garden variety overall similarity relations, "subject to a great deal of indeterminacy."[15]

Garden variety overall similarity relations pay no heed to the distinctions among properties introduced by primitive resemblances. Thus Lewis rightly observes that such relations are indeterminate "(1) as to which respects of similarity and difference are to count at all, (2) as to the relative weights of the respects that do count, (3) as to the minimum standard of similarity that is required, and (4) as to the extent to which we eliminate candidates that are similar enough when they are beaten by competitors with stronger claims." But overall perfectly natural similarity heeds the differences in properties which primitive resemblances introduce. Consequently, it is not indeterminate in any of ways Lewis lists: only perfectly natural similarities count, all perfectly natural similarities count equally, no less-than-perfectly natural similarities count at all, and a candidate is eliminated only when it shares fewer perfectly natural properties than another. Of course, overall perfectly natural similarity admits of ties—as in a case where *y* has exactly three perfectly natural properties, and both *x* and *z* share all three. In such cases I prefer to say that both *x* and *z* are overall perfectly naturally similar to *y*. The important point is that ties do not amount to indeterminacy. Both or neither *x* and *z* are overall perfectly naturally similar to y independently of how we conceive the relata.

---

[15] Lewis, "Postscript to 'Counterpart Theory and Quantified Modal Logic'," p. 42.

My defense of the first premise in my overall argument is now complete. I have shown that if there are natural properties in the counterpart theorist's ontology, then a determinate relation of comparative overall natural similarity obtains between any number of objects independently of the way they are conceived, described, or referred to. As my second premise predicts, and Lewis grants, real essences follow. For if there are overall perfectly natural similarities, then there are maximally natural counterparts of things:

> x is a *maximally natural counterpart* of $y =_{df} x$ and y inhabit different possible worlds and compared to everything else in y's world x is overall perfectly naturally similar to y.

And if there are maximally natural counterparts of things, then things have maximally natural, or real, essences.

> the *real essence* of $x =_{df}$ the class of properties, P_1–P_n, containing all and only the properties shared by all of x's maximally natural counterparts.

In sum, Lewis's natural properties give us natural similarities, which give us perfectly natural similarities, which give us overall perfectly natural similarities, which give us maximally natural counterparts, which give us real essences.

III

The definitions discussed above chart a straight path from the natural properties Lewis claims his ontology needs to the essentialist consequences he disavows. I conclude by answering the objections I anticipate. I see only two ways to resist my argument. One strategy is to deny that the consequences I have drawn from the addition of natural properties to counterpart theory follow; the other strategy is to grant that the consequences I have drawn do indeed follow, but to deny that they amount to essentialism.

The first strategy fails for two reasons. First, the objector would have to show that it is possible for there to be natural properties but no overall perfectly natural similarities—that is, possible for the antecedent of my first premise to be true, while the consequent is false. But the relation of overall perfectly natural similarity is defined in such a way that it is entailed by the notion of naturalness Lewis grants, and nothing more. Thus the objector would have to abuse either Lewis's terms or mine—just as one who denies that the existence of unmarried males implies the existence of bachelors abuses one term or the other.

Second, a counterpart theorist may claim that Lewis has granted too much by way of naturalness, and that a less robust account of naturalness could secure the benefits of Lewis's notion, without generating the consequences I have drawn. The burden is on the objector to offer an alternative account of naturalness. As far as I know, no such alternative is on offer. Further, there is good reason to doubt that an account of naturalness suited to the objector's purposes will be forthcoming.

My argument can accommodate a much weaker notion of naturalness than Lewis grants, and, more importantly, the weakest notion of naturalness suited to the needs of counterpart theory. Specifically, my argument can do without the perfectly natural properties Lewis defines. All my argument requires is that there be a fact of the matter about where on the hierarchy of naturalness the various properties of objects fall. Counterpart theorists will have to grant this much. For in the absence of objective distinctions among degrees of naturalness, natural properties are too abundant to provide a sufficiently inegalitarian theory of properties.

To offer an account of naturalness which admits only of various degrees of naturalness but not perfect naturalness, counterpart theorists will have to deny what I assumed (for the sake of simplicity) at the outset: that naturalness is introduced into the counterpart theoretic ontology by primitive resemblance relations. For this supposition, coupled with Lewis's constraints on the composition of classes, entails perfectly natural properties. But suppose naturalness consists instead in a primitive property of classes. On this supposition, counterpart theorists are free to suppose, as they must, that the primitive property of classes comes in various degrees. But they are also free to suppose that this primitive feature of classes does not admit of a maximum. They need not suppose any class is so natural that there is not another which is more natural.

Without perfectly natural properties, the connection my definitions forge between natural properties and overall perfectly natural similarities is broken. Thus maximally natural counterparts and essences, as defined above, would not follow from the introduction of natural properties. But the essentialist consequences would follow by slightly modified definitions. Rather than defining my similarity relations in terms of perfectly natural property sharing, I define equally unambiguous relations in terms of sharing a thing's most natural property, and the longest consecutive series of a thing's most natural properties. On the revised account of naturalness, there remains a determinate fact of the matter about which of an object's properties are its most natural, its second most natural, its third, and so on. Given the counterpart theorist's principles for the composition of classes, there also remains a fact of the matter about which objects share which properties.

My definitions may thus be reconstructed as follows:

> x is *maximally natural similar* to y $=_{\text{df}}$ x shares y's most natural property ($=_{\text{df}}$ x is a co-member of the most natural class to which y belongs).
>
> x is *overall maximally naturally similar* to y $=_{\text{df}}$ x shares y's most natural property and there is no thing, z, distinct from x and y, such that z is perfectly naturally similar to y, and z, rather than x, shares the longest consecutive series of y's most natural properties, starting with y's most natural property.

As with the earlier definitions, the relations defined above admits ties. For y may have more than one property on any wrung in the hierarchy of natural properties. Consequently, there may be more than one way to achieve first, second, or *n*th degree natural similarity to y. Also as before, things which do not share y's most natural property may achieve an extraordinary degree of similarity to y, but not maximally natural similarity. Nothing can be maximally naturally similar to y which does not share y's most natural property, and nothing can be overall maximally naturally similar to y which does not share a longer consecutive series than any competitors of y's most natural properties, beginning with y's most natural property. But most importantly, as before, maximally natural similarity relations are not vague. It is not indeterminate as to which respects of similarity count, which count most, what minimum standard of similarity is required, or at what point candidates are eliminated. Therefore determinate counterpart relations and essences may be defined in terms of overall maximally natural similarities, just as they were in terms of perfectly natural similarities.

In short, any definition of naturalness which provides for objective distinctions among degrees of naturalness is going to have the same essentialist consequences as the definition Lewis offers. Any definition which does not provide for objective distinctions among degrees of naturalness is not going to earn its keep. For this reason, I doubt that an account of naturalness suited to the needs of counterpart theory but lacking essentialist consequences will be forthcoming.

The second strategy for resisting my argument is to deny that the consequences I have drawn from adding natural properties to counterpart theory amount to essentialism. For counterpart theory to be essentialist, recall, it must ascribe determinate modal properties to some objects (or, equivalently, determinate truth values to some *de re* modal propositions) independently of the way the objects are conceived, described, or referred to. The reason to think counterpart theory is not essentialist is that it ascribes only vague modal properties to objects. The ontology contains counterparts which constitute any *x*'s being possibly *F* and not possibly *F*, leaving it to context to specify which counterparts any modal claim is to be evaluated against. So it is not the ontology but context that ascribes determinate modal properties to objects.

The second objection contends that my definitions do not change this situation. The objection notes that adding natural properties to the counterpart theorist's ontology eliminates nothing from it. Even granting that the ontology with natural properties contains perfectly natural counterparts, the ontology still contains a "countless throng" of less-than-maximally natural properties, and therefore less-than-maximally natural similarity relations, counterparts and essences. According to this objection, it is therefore still up to context to determine whether modal claims about any *x* are to be evaluated against the natural or unnatural modal facts.

The objection is certainly right that the "countless throng" remains. It is also right that we may choose to talk about things in more or less unnatural terms, and by dropping certain contextual clues we may indicate precisely how we want our modal claims to be evaluated. But neither point shows that the consequences I have drawn fall short of essentialism.

To succeed my argument need not eliminate all but one counterpart relation for any object. If this is what the second objection amounts to, then it misunderstands what the naturalness of real essences entails. For my argument is just as successful if it shows that one counterpart relation stands out from the crowd for the purposes of characterizing the modal properties of an object. The anti-essentialism of counterpart theory *without* natural properties is not just a consequence of the fact that the ontology contains so many counterpart relations for each object. It is a consequence of the abundance of counterpart relations together with the fact that the ontology does nothing to privilege one counterpart relation over others. Counterpart theory *with* natural properties contains just as many counterpart relations, but does not suppose they are all on par, ontologically speaking. Some are more natural, and therefore play a more significant role in a complete characterization of the world. Some are maximally natural, and therefore play the most significant role. It is not context, but the facts about objective sameness and difference, that determines which counterpart relations matter most metaphysically.

To succeed my argument need not show that, independent of context, ordinary *de re* modal claims are about objects under their most natural counterpart relations. If this is what the second objection amounts to, it confuses metaphysics with semantics (or, better still, pragmatics). The issue with regard to essentialism is not indeterminacy of modal expression, but of modal fact. Context determines precisely which modal proposition a *de re* modal statement expresses because context determines precisely what proposition any claim expresses. The metaphysical issue is whether the ontology privileges one way of disambiguating the counterpart relation itself—not our claims about it. My argument shows that there are maximally natural counterpart relations "out there" in the counterpart theoretic ontology for us to talk about and to evaluate our *de re* modal claims against. This point holds whether or not we opt to speak about the most natural modal facts, whether we succeed in speaking about them when we try, or even whether we are in the position to know the difference.

While my argument succeeds without showing that any particular modal claims express the most natural modal facts, it is also worth noting there is very good reason to suppose that the most natural modal facts provide the best interpretation ordinary modal discourse. Part of the reason Lewis introduces natural properties into his ontology is to equip counterpart theory to account for the content of language and thought. The most eligible overall interpretation of our language, according to Lewis, is the most natural.[16] So the reason to think ordinary modal discourse is about objects under their maximally natural counterpart relations is that these relations are maximally natural. In the same way, the reason to think ordinary discourse about blue things is about blue rather than grue things is that the class with which blue things is identified is the more natural class. The assumption that ordinary modal discourse is about maximally natural modal facts is thus well grounded. This is not to deny that contextual pressures may bring it about that modal claims are about things under less-than-maximally natural counterpart relations. All these relations, and the modal facts they constitute, remain part of the ontology. They are "out there" for us to talk about. But this is no reason to think that ordinary modal claims are vague, any more than the existence of the class with which grue objects is identified provides a reason to think ordinary claims about blue objects are vague.

The two strategies I have considered for rejecting my conclusion both fail. They provide no reason to doubt either that maximally

---

[16] "Putnam's Paradox," pp. 65–66.

BOOK REVIEWS

The Reasons of Love. HARRY G. FRANKFURT. Princeton: Princeton University Press, 2004. 110 p. Cloth $19.95.

Harry Frankfurt's *The Reasons of Love* is about love in the way in which his bestseller, *On Bullshit*,[1] is about that. It takes a phenomenon that, while engaging enough, seems to lie on the periphery of serious philosophical interest, and it then proceeds to show that the phenomenon is, in fact, key to a question of recognized importance. Just as we learn from bullshit something about the value of truth, so too we learn from love something about how we should live. When we ask how we should live, is the answer there for us to discover by reasoning? Or is it left to us to create by desire or choice? Can we be right or wrong in what we are most deeply committed to? Or is it nonsense to ask, since what we are most deeply committed to sets the standard of right and wrong? Like many contemporary philosophers, Frankfurt takes the latter view: that one's normative reasons arise from one's own psychology. "[T]he most basic and essential question for a person to raise concerning the conduct of his life cannot be the *normative* question of how he *should* live," Frankfurt writes, "That question can sensibly be asked only on the basis of a prior answer to the *factual* question of what he actually *does* care about" (26), and indeed, as we learn, only on the basis of a prior answer to the factual question of what he loves. But Frankfurt's version of the view that our reasons arise from our psychology is, in several ways, quite unlike other versions that one encounters in the literature.

To begin with, he articulates the view with an unusually discriminating account of what our psychology contains. Cautioning against the obscuring promiscuity of words such as "want" and "desire," in which the view is so often exclusively formulated, Frankfurt commends us to distinguish the phenomenon of caring, along with one of its consequences, importance, and one of its species, love.

Caring may involve a structure of desire, but it is a special structure. To care about something is not simply to desire it, but also to desire that one desire it: to desire that one's desire for it continue and prevail. Moreover, this must be "a desire with which the person

---

[1] Princeton: University Press, 2005.

0022-362X/06/0301/43–50

identifies himself, and which he accepts as expressing what he really wants" (16).[2]

If, for Frankfurt, there is a general answer to the question, "What should I do?" it would seem to be: "What it is most important to you to do." But this is consistent with his view that caring is the source of reasons, for caring about something is what makes it important to one. While something may be important to one even though one does not care about it, this is only in virtue of its affecting something else that one does care about.

Three main differentia distinguish love from the genus of caring. First, love is disinterested. Whereas one can care about something merely as a means to something else that one cares about, one cannot love something as a means. Second, love is particular. Whereas one can care about something merely as a member of a class, one cannot love something as a member of a class. A humanitarian may care about my beloved in virtue of her being a human being, but I love my beloved simply in virtue of her being *her*. Finally, love is not voluntary. Whereas one may have more or less direct voluntary control over caring about certain things, one does not have such control over loving it. From this analysis of love in general, Frankfurt arrives at two characteristically startling conclusions about love of oneself in particular. These are that far from being a perversion of love, self-love is its purest form, and that self-love represents a kind of primitive aversion to boredom: a concern to be interested in something.

It is not simply the internal richness of Frankfurt's picture of the valuing mind that is distinctive, but also the density of its ties to other philosophical questions. In passing, Frankfurt connects his account of caring to the work for which he is perhaps best known, on personhood and freedom of the will. By caring about things, he suggests, we define ourselves; by identifying with some desires and standing opposed to others, we shape a person from what would otherwise be a welter of brutely conflicting desires. What most clearly distinguishes persons from mere animals, then, is not rationality or intelligence, but instead reflexivity: the awareness of, and hence the ability to take a stake in, one's own mental life. Furthermore, by caring about things, we make it possible for us to be free or unfree: to succeed or fail to be guided by that with which we identify ourselves. These glimpses of meshing gears are deeply satisfying, and they confirm the reader's sense of a fully wrought system underlying the book's pleasingly brisk and anecdotal style.

---

[2] Frankfurt says little in *The Reasons of Love* to explain how this identification might be brought about, or what it might come to. In "Identification and Externality," in *The Importance of What We Care About* (New York: Cambridge, 1988), pp. 58–68, Frankfurt argued that such identification could not result from, or amount to, the presence of some higher-order attitude. One could always ask whether the agent was identified with this higher-order attitude, and the answer could be "yes" only if the agent had an attitude of an even higher order toward it, provoking an infinite regress. Instead, Frankfurt then suggested, identification results from, or is constituted by, a higher-order decision regarding one's attitudes (68 note 3). However, he no longer appears to understand identification in this way. In "Reply to Richard Moran," in Sarah Buss and Lee Overton, eds., *Contours of Agency: Essays on Themes from Harry Frankfurt* (Cambridge: MIT, 2002), pp. 218–26, Frankfurt contends that the same regress threatens the attempt to account for identification in terms of "any... kind of activity" (220). Moreover, no decision could establish the kind of identification necessary for love, since love is not under our voluntary control. Instead, as Frankfurt suggests in "The Faintest Passion," in *Necessity, Volition, and Love* (New York: Cambridge, 1999), pp. 95–107, identification with a desire is constituted by a purely negative condition: the absence of any psychic tendency opposed to that desire.

The final distinguishing feature of Frankfurt's version of the view that our reasons arise from our own psychology lies in his grounds for it. In the literature, one finds the view most often fueled by motivational, metaphysical, or epistemological concerns. How could the making of, or the truth of, a judgment about what one ought to do motivate in the way that it does, unless it somehow depended on our psychology? What kind of fact could it be that one ought to do something if not a fact about one's psychology? How else could we ever come to know anything about it? Frankfurt may share something of these last epistemological concerns. What would an argument for the ultimate value of something look like, he asks, and how could it hope to get off of the ground (23–26)? But he also seems drawn to the view that reasons arise from our own psychology by a far less abstract conviction: that the view is straightforwardly confirmed by our lived experience of love.

Love, Frankfurt contends, does not answer to reasons. It is not a response to the perceived inherent value of its object: a response that would be correct insofar as its object had that value, and incorrect insofar as its object lacked it.[3] This is not cheap romanticism, but the seemingly irresistible consequence of a few commonplaces. I may love something, while you do not, even though we agree about its intrinsic properties. Moreover, I may love one thing, but not another, even though I believe that both have the same intrinsic properties. Our love for our children precedes any beliefs about their intrinsic properties, and once in place, is insensitive to changes in those beliefs.

---

[3] This does not mean that Frankfurt does or must deny that what we love may have or lack value independently of our loving or not loving it, or, more generally, that things can have or lack value independently of our psychological orientation toward them. All he is committed to, it seems, is the claim that if there are such independent values, our reasons do not depend on them.

Frankfurt does not seriously consider the idea that we might be mistaken in loving as we do: that either I ought to stop loving what I do, or you ought to start; that I ought to love substitutes; that I ought to wait to see how my child is before loving it and to withdraw my love should it disappoint me. It goes without saying that the idea is preposterous. Indeed, it may not even be coherent; to love like that would be not to love at all.

It would be no less preposterous, however, to claim that we are mistaken in taking what we love to be relevant to how we should live. Just as it cannot be claimed that reasons give rise to love, it cannot be denied that love gives rise to reasons. There are reasons of love, as the title announces, but these are not the reasons to which love responds, but instead the reasons that love creates. Similar observations suggest that such reasons arise not simply from love for persons, but also from love for groups, traditions, callings, causes, and many other particular objects of devotion.

This argument from love is not only constructed from more concrete materials than the motivational, metaphysical, and epistemological arguments alluded to earlier. It also has different consequences. First, it lacks the reductive tendency of at least some of those arguments. For example, the metaphysical concern, about what the fact that one "ought" to do something could even be, seems to demand not simply that the fact that one "ought" to do something be explained by a fact about one's psychology, but moreover that it be nothing more than a fact about one's psychology. There is nothing in the argument from love that presses for this stronger claim, however, and there is no clear indication that Frankfurt accepts it.

Second, the argument from love seems, at least at first glance, not to support as general a thesis about the source of normativity. On its own, the argument seems to show only that some of our reasons are rooted in our psychology, not that all of our reasons are rooted in our psychology. For example, it appears compatible with Frankfurt's observations about love as it is ordinarily understood—love, say, for people, callings, and nations—that one ought to comply with the dictates of morality whether or not one cares to.

This restricted conclusion would still be quite congenial to Frankfurt's guiding skepticism about the reach of reason into our practical lives. In the event, however, Frankfurt asserts something broader. Not simply are there reasons of love, he proposes, but all reasons are reasons of love (55–57). One wonders whether this is, strictly speaking, Frankfurt's position. For he says nothing elsewhere to rule out reasons that arise, at the most fundamental level, from caring that, unlike love, is focused not on particulars as such, but instead on members of a class. Indeed, the reasons of morality, on Frankfurt's understanding, might well be reasons of this kind. Perhaps it would be more accurate to represent his claim as being that all reasons are, like reasons of love, reasons that arise from caring that is noninstrumental and not under the agent's direct voluntary control.

However this broader claim is interpreted, there is the question of what supports it, if not Frankfurt's observations about love, as it is ordinarily understood. This support may come from the epistemological doubts mentioned earlier. But Frankfurt also seems to suggest, more directly, that just as we readily accept that the reasons involved in love as it is ordinarily understood are rooted in our own wills, so too do we accept, on reflection, that reasons of other kinds are also rooted in our wills. "There may be some individuals for whom a commitment to being morally virtuous is a categorically dominant personal ideal." For such people, he grants, moral requirements will be overriding. "That is not, however, the only intelligible or the only appealing design for a human life. We may find that other ideals and other measures of value attract us, and that they recommend themselves to us forcefully as reasonable competitors for our controlling allegiance" (9). Just as your love of your children may give you reasons to look after them that I do not have, Frankfurt seems to suggest, so too your commitment to a life of sainthood may give you reasons to do your duty that I do not have. Even the authority of morality depends, in the end, on what one cares about.

This broader thesis, that not just some, but all, reasons flow from what one cares about, meets with resistance. We often think that someone ought not to care about what he does, or that he ought to care about something that he does not. Perhaps we do not think this where what he cares about is, say, his own child and the alternative on offer is a stranger's child. Such is the intuitive appeal of Frankfurt's reflections on love, as it is ordinarily understood. But similar thoughts hardly seem unnatural where what one cares about is, say, one's own pointless dissipation, or the destruction of some innocent. It may be optional to see "being morally virtuous" as "a categorically dominant personal ideal." But it does not seem merely optional to give some weight to being minimally morally decent, whether or not one cares about it, or about anything it might serve. Nor does it seem merely optional to care about one's own life, or a great many other things.

In a footnote that appears to address this kind of objection, Frankfurt observes that "we would regard" someone who, say, "cared nothing at all about dying or being mutilated, or about being deprived of all human contact" not merely as "atypical," but also as "irrational," "crazy," and "lacking reason" (45 note 5). Since this seems to be normative criticism of what a person fails to care about, it seems to be at odds with Frankfurt's insistence that caring constitutes the highest court of appeal on normative questions. The resolution of this tension, it appears, lies in fact that the irrationality is being imputed from our point of view.[4] Since we care so much about our lives, bodily integrity, and commerce with our fellows, the practical possibilities open to him are, for us, "unthinkable"; we cannot "bring ourselves to do or to accept" them. This is the sense in which they appear to us as "irrational." This does not mean, however, that there is some objective sense in which such a person would be wrong, rather than simply not like us.

The fact that we criticize others for what they care about, as we have just seen, seems, at least at first, to tell against Frankfurt's thesis that caring is the source of normativity. And he appears to respond, again as we have just seen, by interpreting that criticism as itself an expression of caring. The fact that we sometimes ask, in our own case, whether we ought to care about what we do also seems to tell against Frankfurt's thesis. And he appears to respond in a similar way, interpreting such questioning as itself expressing a desire not to care for what is under review. "Coping with our troubled and restless uncertainty about how to live does not require us to discover what way of living can be justified by definitive argument. Rather, it requires us simply to understand what it is that we ourselves really care about, and to be decisively and robustly confident in caring about it" (28). This "confidence" is not a belief that one has good grounds for caring about what one does, or that one is correct in so caring. If I understand it correctly, it is not, as the word might suggest, a belief at all. It instead seems closer to what Frankfurt elsewhere calls "wholeheartedness": the absence of any desire opposed to caring about the things one cares about (cf. 95 note 7). Since normative questions are settled by what we care about, Frankfurt's thought seems to be, only such a countervailing desire could provoke serious doubt about whether we ought to care about what we do. When combined with Frankfurt's later suggestion that to love oneself is to care about what one cares about (95), this diagnosis takes on a kind of *ad hominem* sting. Only a kind of self-hatred, or at least lack of self-love, the reader is left to conclude, could drive a person to question whether he ought to care about what he does.

All of this is ingenious and marvelously consistent. For those who have been won over to Frankfurt's view, it offers a new way of understanding the opposition. But one wonders how much it will do to convince those who are likely to bring up the problem. To someone who seriously asks what he should care about, the exhortation to be confident is liable to seem no answer at all. Perhaps he is able to be confident, in Frankfurt's sense: to care about something, or not to care about it, without residue. His question, as he understands it, is whether he would be right in so caring. The enduring intelligibility of this question, he will argue, is a datum counting against our acceptance of Frankfurt's theory.

---

[4] Compare "Rationality and the Unthinkable," in *The Importance of What We Care About*, p. 177–90.

So far I have been granting that Frankfurt shows that the reasons involved in love as it is ordinarily understood—love for, say, our child, life's work, or people—arise from our psychology. I have been asking whether this is true of reasons of other kinds, such as those of morality. But is it really so obvious that the reasons involved in love as it is ordinarily understood arise from our psychology?

What if someone, peering into the schoolyard, found himself caring about a stranger's child rather than his own child, Alice? Would we not think that he ought to care about something other than he does? Perhaps this parent's affections would be "unthinkable" for us, in the way that a normal parent's would not be. But this itself seems to tell against Frankfurt's view. What we would be finding unthinkable in this case is not failing to love Alice: a person viewed simply as a particular. After all, we too, quite thinkably, fail to love Alice. What we would be finding unthinkable is failing to love one's child: failing to love a person viewed as falling under a certain description. This seems to conflict not only with Frankfurt's analysis of love, which insists that it attaches to particulars as such, but also with his case that there are no reasons for love. For it offers us a possible reason for loving a person: that one shares a history of a certain kind with her, such as being her parent.

This is something that might be a reason for loving a person: it is a generalizable property of a person that might render love for her intelligible and appropriate. And since this property is relational, rather than intrinsic, the proposal that it represents a reason for love is untouched by the commonplaces about love mentioned earlier. Indeed, the proposal positively predicts that I will love my daughter, whereas you may not, even though we agree about her intrinsic properties; that I will love my daughter, but not any *Doppelgänger* that might appear, even one with plainly the same intrinsic properties (whereas I would love a second daughter with vastly different intrinsic properties); and that my love for my daughter will precede any beliefs about her intrinsic properties and, once in place, be insensitive to changes in those beliefs. In sum, Frankfurt's argument that there can be no reasons for loving a person or thing depends on the implicit assumption that any such reasons would have to be intrinsic properties of that person or thing. Once it is seen that such reasons might be extrinsic properties, such as historical relations to the lover of a certain type, the argument is no longer so clearly decisive (whatever force Frankfurt's other arguments, such his epistemological doubts about settling normative questions in general, might retain).

Inclined though I am to resist some of its conclusions, this is an undeniably brilliant, resourceful, and provocative book. Still more, these doubts are accompanied by several points of agreement. Perhaps the most important of these points of agreement is with Frankfurt's manifest conviction that a theory of practical reason must attend faithfully to its most intimate and familiar exercises.

NIKO KOLODNY

University of California, Berkeley

MEMO TO AUTHORS

The Journal of Philosophy is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally publish to no more than 7,500 words, or 18 *Journal* pages. A more precise way of determining length is by characters (including spaces): 2,850 characters per *Journal* page, or 51,300 characters per *Journal* article of 18 pages.

The editors request that three copies (or an electronic version) of all manuscripts submitted be prepared according to the following guidelines:

(1) All material, including the quotations and footnotes, must be typed, double-spaced, and one-sided, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into footnotes.
(3) Footnotes should be complete, according to *Journal* style (see any issue). Journal citations should include: name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include: publisher, U.S. city of publication, and date. Footnotes should be gathered at the end of the article.
(4) Quotations should be checked by the author; capitalization, italics, spelling, and the like should match the original.
(5) If there are diagrams, they should be provided camera-ready by the author. If there are special symbols, they should be widely spaced and clearly marked. Standard logic symbols are available, and letters of the Greek and script alphabet.
(6) No underlining; please limit use of italics to technical terms (when they are introduced) and to variables.
(7) No contractions.

The editors are willing to read and evaluate a manuscript that does not meet these guidelines, provided the author agrees to meet them before submitting a final manuscript. The editors may ask for revision of excessively long manuscripts, and may require reformatting of accepted manuscripts. They also request an electronic version with the final hard copy.

The Journal of Philosophy does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere. While we do not publish manuscripts that mainly comment on articles published elsewhere, we do have a "Comments and Criticism" section in which short responses to *Journal* articles may be published.

Manuscripts may be sent to: John Smylie, Managing Editor, *The Journal of Philosophy*, 709 Philosophy Hall, 1150 Amsterdam Avenue, Mail Code 4972, New York, NY 10027, Email jys3@columbia.edu.

NEW BOOKS

RESCHER, NICHOLAS: *Common Sense: A New Look at an Old Philosophical Tradition.* Milwaukee, WN: Marquette, 2005. 270 p. Cloth $20.00; *Reason and Reality: Realism and Idealism in Pragmatic Perspective.* Lanham, MD: Rowman and Littlefield, 2005. ix, 169 p. Cloth $60.00; *Scholastica Meditations.* Washington, DC: Catholic University of America, 2005. x, 169 p. Cloth $49.95.

ROCKWELL, W. TEED: *Neither Brain nor Ghost: A Nondualist Alternative to the Mind-Brain Identity Theory.* Cambridge: MIT, 2005. xvii, 231 p. Cloth $36.00.

ROLL-HANSEN, NILS: *The Lysenko Effect: The Politics of Science.* Amherst, NY: Prometheus, 2005. 335 p. Paper $25.00.

RORTY, RICHARD, and GIANNI VATTIMO: *The Future of Religion.* SANTIAGO ZABALA, ed. New York: Columbia, 2005. x, 91 p. Cloth $24.50.

ROSENBERG, GREGG: *A Place for Consciousness: Probing the Deep Structure of the Natural World.* New York: Oxford, 2004. xviii, 325 p. Cloth $45.00.

ROSENTHAL, EDWARD C.: *The Era of Choice: The Ability to Choose and Its Transformation of Contemporary Life.* Cambridge: MIT, 2005. xii, 331 p. Cloth $32.50.

ROSENZWEIG, FRANZ: *The Star of Redemption.* Madison: Wisconsin UP, 2005. xx, 459 p. Cloth $65.00, paper $19.95.

ROSLER, ANDRES: *Political Authority and Obligation in Aristotle.* New York: Oxford, 2005. xiv, 298 p. Cloth $74.00.

RUALET, GÉRARD: *Critical Cosmology: On Nations and Globalization: A Philosophical Essay.* Lanham, MD: Lexington, 2005. xix, 99 p. Cloth $58.00, paper $19.99.

RUSE, MICHAEL: *The Evolution-Creation Struggle.* Cambridge: Harvard, 2005. 327 p. Cloth $29.95.

RUTHERFORD, DONALD, and J.A. COVER: *Leibniz: Nature and Freedom.* New York: Oxford, 2005. xvi, 265 p. Cloth $74.00, paper $24.95.

RYCKMAN, THOMAS: *The Reign of Relativity: Philosophy in Physics, 1915–1925.* New York: Oxford, 2005. ix, 317 p. Cloth $60.00.

RYNHOLD, DANIEL: *Two Models of Jewish Philosophy: Justifying One's Practices.* New York: Oxford, 2005. 262 p. Cloth $95.00.

SARKAR, SAHORTA: *Molecular Models of Life: Philosophical Papers on Molecular Biology.* Cambridge: MIT, 2005. xvi, 396 p. Cloth $38.00.

SAVILE, ANTHONY: *Kant's Critique of Pure Reason: An Orientation to the Central Theme.* Malden, MA: Blackwell, 2005. 148 p. Cloth $54.95, paper $21.95.

SCHELLENBERG, J.L.: *Prolegomena to a Philosophy of Religion.* Ithaca, NY: Cornell, 2005. xiii, 226 p. Cloth $45.00.

SCHROEDER, TIMOTHY: *Three Faces of Desire.* New York: Oxford, 2004. viii, 213 p. Cloth $35.00.

SCHWEIG, GRAHAM M.: *Dance of Divine Love: India's Classic Sacred Love Story: The Rāsa Līlā of Krishna.* Princeton: University Press, 2005. xxiii, 390 p. Cloth $35.00.

SEESKIN, KENNETH: *Maimonides on the Origin of the World.* New York: Cambridge, 2005. viii, 215 p. Cloth $55.00.

SEHON, SCOTT: *Teleological Realism: Mind, Agency, and Explanation.* Cambridge, MA: MIT Press, 2005. xii, 245 p. Cloth $36.00.

SEUNG, T.K.: *Nietzsche's Epic of the Soul: Thus Spoke Zarathustra.* New York: Lexington, 2005. xxvii, 371p. Cloth $83.00, Paper $27.95.

SHELBY, TOMMIE: *We Who Are Dark: The Philosophical Foundations of Black Solidarity.* Cambridge: Harvard, 2005. xiv, 320 p. Cloth $27.95.

SHERMAN, NANCY: *Stoic Warriors: The Ancient Philosophy Behind the Military Mind.* New York: Oxford, 2005. xii, 242 p. Cloth $26.00.

SIDER, DAVID: *The Library of the Villa Dei Papiri at Herculaneum.* Los Angeles: Getty, 2005. viii, 123 p. Cloth $40.00.

SIMON, ROGER I.: *The Touch of the Past: Remembrance, Learning, and Ethics.* New York: Palgrave Macmillan, 2005. xi, 192 p. Cloth $75.00, paper $25.95.

SINGER, IRVING: *Sex: A Philosophical Primer, Expanded Edition.* Lanham, MD: Rowman and Littlefield, 2004. xxxvi, 134 p. Cloth $55.00, paper $18.95.

SMALL, ROBIN: *Nietzsche and Rée: A Star Friendship.* New York: Oxford, 2005. xxiv, 247 p. Cloth $45.00.

SMITH, ANDREW: *Philosophy in Late Antiquity.* New York: Routledge, 2004. xi, 151 p. Cloth $99.95, paper $29.95.

SOLOMON, ROBERT C.: *In Defense of Sentimentality.* New York: Oxford, 2004. xvii, 297 p. Cloth $49.95.

SORELL, TOM: *Descartes Reinvented.* New York: Cambridge, 2005. xxii, 180 p. Cloth $75.00.

SPARKE, MATTHEW: *In The Space of Theory: Postfoundational Geographies of the Nation-State.* Minneapolis: Minnesota UP, 2005. xxxv, 395 p. Cloth $77.95, paper $25.95.

SPENCER, MARK G.: *David Hume and Eighteenth-Century America.* Rochester, NY: University Press, 2005. xii, 534 p. Cloth $90.00.

STEINER, GARY: *Anthropocentrism and Its Discontents: The Moral Status of Animals in the History of Western Philosophy.* Pittsburgh: University Press, 2005. 344 p. Cloth $37.50.

STERN, DAVID G.: *Wittgenstein's "Philosophical Investigations": An Introduction.* New York: Cambridge, 2004. xv, 208 p. Cloth $65.00, paper $23.99.

STERN, LAURENT: *Interpretive Reasoning.* Ithaca, NY: Cornell, 2005. xiv, 214 p. Cloth $39.95.

STOUT, JEFFREY: *Democracy and Tradition.* Princeton: University Press, 2004. xv, 349 p. Cloth $49.95, paper $19.95.

SZABÓ, ZOLTÁN GENDLER: *Semantics versus Pragmatics.* New York: Oxford, 2005. viii, 465 p. Cloth $115.00, paper $45.00.

TALBOTT, WILLIAM J.: *Which Rights Should Be Universal?* New York: Oxford, 2005. viii, 222 p. Cloth $35.00.

TANCREDI, LAURENCE: *Hardwired Behaviour: What Neuroscience Reveals about Morality.* New York: Cambridge, 2005. xiv, 226 p. Cloth $28.99.

TAYLOR, PETER J.: *Unruly Complexity: Ecology, Interpretation, Engagement.* Chicago: University Press, 2005. xxi, 289 p. Paper $25.00.

THAGARD, PAUL: *Mind: Introduction to the Cognitive Science, 2nd Edition.* Cambridge: MIT, 2005. x, 266 p. Paper $30.00.

THOMSON, ALEX: *Deconstruction and Democracy.* New York: Continuum, 2005. x, 226 p. Cloth $105.00.

THOMSON, IAIN D.: *Heidegger on Ontotheology: Technology and the Politics of Education.* New York: Cambridge, 2005. xx, 202 p. Cloth $70.00, paper $24.99.

THOMSON, KEITH. *Before Darwin: Reconciling God and Nature.* New Haven: Yale, 2005. xiv, 314 p. Cloth $25.00.

TOPPER, KEITH: *The Disorder of Political Inquiry.* Cambridge: Harvard, 2005. xi, 322 p. Cloth $45.00.

UTO, SHOKICHI: *Constructics: A Methodology of Theory Construction.* Lanham, MA: University Press of America, 2005. xxii, 283 p. Paper $41.00.

VAN LAMBALGEN, MICHIEL, and FRITZ HAMM: *The Proper Treatment of Events: Explorations in Semantics.* Malden, MA: Blackwell, 2005. xii, 251 p. Cloth $74.95, paper $39.95.

VANDEVELDE, POL: *The Task of the Interpreter: Text, Meaning, and Negotiation.* Pittsburgh: University Press. x, 240 p. Cloth $55.00, paper $22.95.

VAUGHT, CARL G.: *Metaphor, Analogy, and the Place of Places.* Waco, TX: Baylor, 2004. xv, 206 p. Cloth $29.95.

WALL, JOHN: *Moral Creativity: Paul Ricoeur and the Poetics of Possibility.* New York: Oxford, 2005. viii, 234 p. Cloth $55.00.

WALLACE, ROBERT M.: *Hegel's Philosophy of Reality, Freedom and God.* New York: Cambridge, 2005. xxxiii, 346 p. Cloth $75.00.

WARD, STEPHEN J.A.: *The Invention of Journalism Ethics: The Path to Objectivity and Beyond.* London: McGill-Queen's. xii, 360 p. Cloth $65.00.

WARGO, ROBERT J.J.: *The Logic of Nothingness: A Study of Nishida Kitarō.* Honolulu: Hawaii UP, 2005. xi, 241 p. Cloth $45.00, paper $25.00.

WATKINS, ERIC: *Kant and the Metaphysics of Causality.* New York: Cambridge, 2005. xi, 451 p. Cloth $75.00, paper $32.99.

WEBER, SAMUEL: *Targets of Opportunity: On the Militarization of Thinking.* New York: Fordham, 2005. x, 152 p. Cloth $55.00, paper $18.00.

WEIL, SIMONE, and RACHEL BESPALOFF: *War and the Iliad.* New York: Review of Books, 2005. xxiii, 121 p. Paper $14.95.

WEINER, JOAN: *Frege Explained: From Arithmetic to Analytic Philosophy.* Chicago: Open Court, 2004. xv, 179 p. Paper $18.95.

WEINREB, LLOYD: *Legal Reason: The Use of Analogy in Legal Argument.* New York: Cambridge, 2005. viii, 184 p. Cloth $60.00, paper $18.99.

WEIRICH, PAUL: *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances.* New York: Oxford, 2004. x, 267 p. Cloth $55.00.

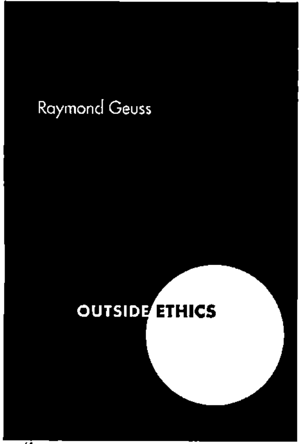
WESTBROOK, ROBERT B.: *Democratic Hope: Pragmatism and the Politics of Truth.* Ithaca, NY: Cornell, 2005. xvi, 246 p. Cloth $29.95.

WESTPHAL, KENNETH R.: *Kant's Transcendental Proof of Realism.* New York: Cambridge, 2005. x, 299 p. Cloth $80.00.

WHEELER, MICHAEL: *Reconstructing the Cognitive World: The Next Step.* Cambridge: MIT, 2005. xi, 340 p. Cloth $35.00.

WHITE, MORTON: *From a Philosophical Point of View.* Princeton: University Press, 2005. viii, 355 p. Cloth $45.00.

Important



OUTSIDE ETHICS

Raymond Geuss

Seeking to expand the scope of contemporary moral and political philosophy, Raymond Geuss presents essays bound by a shared skepticism about a number of common assumptions that are widely shared in the modern world. These include the assumption that politics is best understood as a kind of applied ethics and that the most important feature of any person's life is the way in which that person answers the question 'What ought I to do?'. In moving from issues such as freedom, virtue, the good life, and happiness to the role of history, aesthetic concerns, and the incompleteness of human life, the essays attempt to defend the importance of historically informed philosophical reflection.

"As a philosophical essayist, Geuss has achieved the height of his art."
—Hans Sluga, author of *Heidegger's Crisis*

Paper $24.95 0-691-12342-X Cloth $59.50 0-691-12341-1 Due January

KIERKEGAARD'S JOURNALS AND NOTEBOOKS

Volume I, Journals AA–DD

Søren Kierkegaard

Bruce H. Kirmmse, Chairman, Editorial Board; Editors, Niels Jørgen Cappelørn, Alastair Hannay, Bruce H. Kirmmse, George Pattison, and Jon Stewart, in cooperation with the Søren Kierkegaard Research Centre, Copenhagen

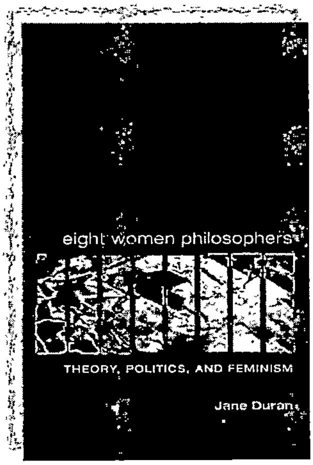
The first of an 11-volume series meticulously produced by the Søren Kierkegaard Research Centre in Copenhagen, this volume represents the first English translation and commentary of Kierkegaard's journals based on up-to-date scholarship. In addition to the early drafts of his published works, the journals contain his thoughts on then-current events and philosophical and theological matters, notes on books he was reading, miscellaneous jottings, and ideas for future literary projects.

Cloth $85.00 0-691-09222-2 Due February

PRINCETON
University Press

800-777-4726
Read excerpts online
www.pup.princeton.edu

UNIVERSITY OF ILLINOIS PRESS



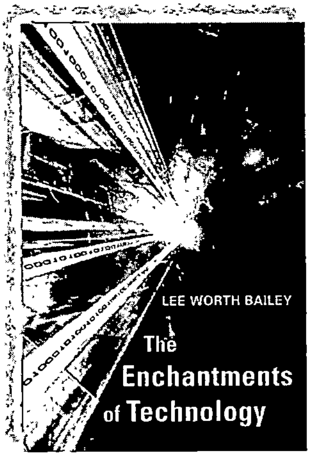
eight women philosophers

THEORY, POLITICS, AND FEMINISM

Jane Duran

This comprehensive overview of eight women philosophers—Hildegard of Bingen, Anne Conway, Mary Astell, Mary Wollstonecraft, Harriet Taylor Mill, Edith Stein, Simone Weil, and Simone de Beauvoir—underscores the continuing significance of these thinkers for contemporary scholars.

Cloth, $65.00; Paper, $29.95



the enchantments of technology

Lee Worth Bailey

Erases the conventional distinction between myth and machine in order to explore the passionate foundations concealed in technological culture and address its complex ethical, moral and social implications.

Cloth, $50.00; Paper, $25.00

From your bookseller, call 800-537-5487, or visit www.press.uillinois.edu

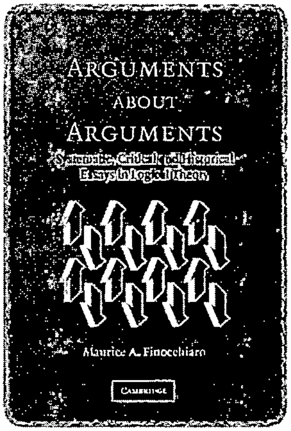
CAMBRIDGE

OUTSTANDING SCHOLARSHIP

ARGUMENTS ABOUT ARGUMENTS
Systematic, Critical, and Historical Essays in Logical Theory
Maurice A. Finocchiaro

Following an approach that is empirical but not psychological, and dialectical but not dialogical, Maurice Finocchiaro defines concepts such as reasoning, argument, argument analysis, critical reasoning, methodological reflection, judgment, critical thinking, and informal logic. Including extended critiques of the views of many contemporary scholars, he also integrates into the discussion Arnauld's Port-Royal Logic, Gramsci's theory of intellectuals, and case studies from the history of science, particularly the work of Galileo, Newton, Huygens, and Lavoisier.

$75.00: Hardback: 0-521-85327-3: 478pp
$31.99: Paperback: 0-521-61853-3


ARGUMENTS
ABOUT
ARGUMENTS
Systematic, Critical, and Historical Essays in Logical Theory
Maurice A. Finocchiaro
CAMBRIDGE

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

You can now subscribe to the *Journal of Philosophy* using

PayPal

To learn more about this easy new option, click on the PayPal icon at **www.jphil.org.**

CAMBRIDGE

OUTSTANDING SCHOLARSHIP

Descartes
A Biography
Desmond Clarke
$40.00: Hardback: 0-521-82301-3: 520pp

The Spiritual Dimension
Religion, Philosophy and Human Value
John Cottingham
$70.00: Hardback: 0-521-84377-4: 198pp
$24.99: Paperback: 0-521-60497-4

Genetics and Christian Ethics
Celia Deane-Drummond
New Studies in Christian Ethics 25
$75.00: Hardback: 0-521-82943-7: 306pp
$29.99: Paperback: 0-521-53637-5

The Cambridge Companion to Kant and Modern Philosophy
Edited by Paul Guyer
Cambridge Companions to Philosophy
$75.00: Hardback: 0-521-82303-X: 736pp
$29.99: Paperback: 0-521-52995-6

The Cambridge History of Eighteenth-Century Philosophy
Edited by Knud Haakonssen
$140.00*: 2-Volume Hardback Set:
0-521-41854-2: c.968pp

From Apology to Utopia
The Structure of International Legal Argument
Martti Koskenniemi
$110.00: Hardback: 0-521-83806-1: 702pp
$50.00: Paperback: 0-521-54696-6

Linguistic Turns in Modern Philosophy
Michael Losonsky
The Evolution of Modern Philosophy
$70.00: Hardback: 0-521-65256-1: 296pp
$24.99: Paperback: 0-521-65470-X

Illegal Beings
Human Clones and the Law
Kerry Lynn Macintosh
$28.00: Hardback: 0-521-85328-1: 286pp

Lawyering Skills and the Legal Process
Caroline Maughan and Julian Webb
Law in Context
$45.00: Paperback: 0-521-61950-5: 468pp

Leo Strauss and the Theologico-Political Problem
Heinrich Meier
Modern European Philosophy
$60.00: Hardback: 0-521-85647-7: 208pp

Aristotle's *Nicomachean Ethics*
An Introduction
Michael Pakaluk
Cambridge Introductions to Key Philosophical Texts
$70.00: Hardback: 0-521-81742-0: 358pp
$27.99: Paperback: 0-521-52068-1

Hardwired Behavior
What Neuroscience Reveals about Morality
Laurence Tancredi
$28.99: Hardback: 0-521-86001-6: 240pp

The Judicial Process
Realism, Pragmatism, Practical Reasoning and Principles
E.W. Thomas
$95.00: Hardback: 0-521-85566-7: 442pp

Phenomenology, Logic, and the Philosophy of Mathematics
Richard Tieszen
$70.00: Hardback: 0-521-83782-0: 368pp

Prices subject to change.

www.cambridge.org/us/philosophy/

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

OUTSTANDING SCHOLARSHIP

The Design Inference
Eliminating Chance Through Small Probabilities
William A. Dembski
$85.00: Hardback: 0-521-62387-1: 272pp
$24.99: Paperback: 0-521-67867-6

The Moral Ecology of Markets
Daniel Finn
$65.00: Hardback: 0-521-86082-2: 184pp
$23.99: Paperback: 0-521-67799-8

Mainstream and Formal Epistemology
Vincent Hendricks
$70.00: Hardback: 0-521-85789-9: 200pp

Aquinas: Summa Theologiae, Questions on God
Edited by Brian Leftow and Brian Davies
Cambridge Texts in the History of Philosophy
$75.00: Hardback: 0-521-82140-1: 344pp
$29.99: Paperback: 0-521-52892-5

Bodies and Souls, or Spirited Bodies?
Nancey Murphy
Current Issues in Theology
$65.00: Hardback: 0-521-85944-1: 164pp
$22.99: Paperback: 0-521-67676-2

Levinas and Theology
Michael Purcell
$70.00: Hardback: 0-521-81325-5: 208pp
$27.99: Paperback: 0-521-01280-5

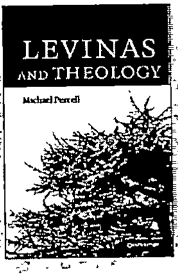
Epistemetrics
Nicholas Rescher
$65.00: Hardback: 0-521-86120-9: 128pp

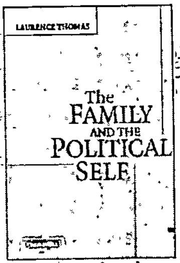
Elements of Justice
David Schmidtz
$70.00: Hardback: 0-521-83164-4: 256pp
$24.99: Paperback: 0-521-53936-6

The Limits of Medicine
Andrew Stark
$70.00: Hardback: 0-521-85631-0: 264pp
$25.99: Paperback: 0-521-67226-0

The Family and the Political Self
Laurence Thomas
$55.00: Hardback: 0-521-85417-2: 200pp
$24.99: Paperback: 0-521-67011-X

Fundamentals of Critical Argumentation
Douglas Walton
Critical Reasoning and Argumentation
$75.00: Hardback: 0-521-82319-6: 360pp
$29.99: Paperback: 0-521-53020-2





Prices subject to change

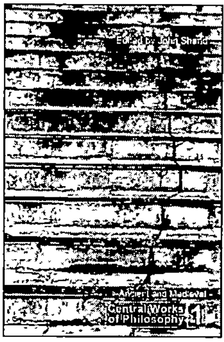
www.cambridge.org/us/philosophy/

CAMBRIDGE
UNIVERSITY PRESS

Central Works of Philosophy
Edited by John Shand

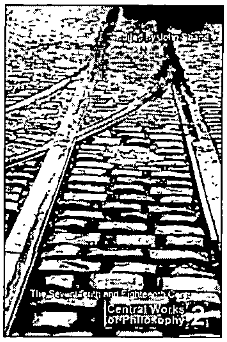
Vol 1: Ancient and Medieval
0-7735-3016-9 $24.95 US / $29.95 CAN paper
0-7735-3015-0 $85.00 cloth
288 pp

"A must-have for any student of philosophy."
James Fieser, University of Tennessee, and editor of the *Internet Encyclopedia of Philosophy.*



Vol 2: Seventeenth and Eighteenth Century
0-7735-3018-5 $24.95 US / $29.95 CAN paper
0-7735-3017-7 $85.00 cloth
256 pp

"A very readable and highly engaging introduction to early modern philosophy." Steven Horst, Wesleyan University


Edited by John Shand
The Seventeenth and Eighteenth Century
Central Works of Philosophy 2

Recently Published

Gilles Deleuze
Key Concepts
Edited by Charles J. Stivale
0-7735-2985-3 $22.95 US / $27.95 CAN paper
0-7735-2984-5 $75.00 cloth
224 pp

"Rigorous and accessible, this book will become the basic starting point for any attempt to teach Deleuze."
Lawrence Grossberg, co-editor of the *Routledge Cultural Studies Reader* and *Cultural Studies*

Jean-Luc Nancy and the Future of Philosophy
B.C. Hutchens
0-7735-2983-7 $22.95 US / $27.95 CDN paper
0-7735-2982-9 $75.00 cloth
224 pp

The first introduction to the work of the contemporary French philosopher.

The Philosophy of Kierkegaard
George Pattison
0-7735-2987-X $22.95 US / $27.95 CAN paper
0-7735-2986-1 $75.00 cloth

"Genuinely original, it shows Pattison's firm mastery of the Kierkegaard corpus."
John Lippitt, University of Hertfordshire and the author of *Humour and Irony in Kierkegaard's Thought*

The Philosophy of Schopenhauer
Dale Jacquette
0-7735-2981-0 $22.95 US / $27.95 CAN paper
0-7735-2980-2 $75.00 cloth

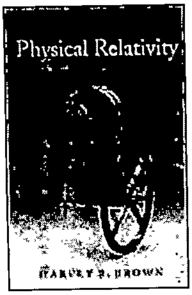
"Jacquette's treatment is both sympathetic and critical and he presents Schopenhauer as a more rigorous thinker than most introductory works do." David Cartwright, University of Wisconsin

McGILL
QUEEN'S
UNIVERSITY
PRESS

For a complete list of our philosophy titles, please visit our website, www.mqup.ca
1-877-864-8477 (Canada)
1-800-666-2211 (US)

New from

Oxford



Philosophical Papers

Volume 1

Peter Unger

Philosophical Papers

Volume 2

Peter Unger

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS

Physical Relativity

Space-time Structure from a Dynamical Perspective

HARVEY R. BROWN

Appearing in the centennial year of Einstein's celebrated paper on special relativity, *Physical Relativity* is an unusual, critical examination of Einstein's thinking that will be of great interest to philosophers of physics, physicists, and historians of science.

2006 $55.00

Ecological Thinking

The Politics of Epistemic Location

LORRAINE CODE

Drawing on ecological theory and practice, on naturalized epistemology, and on feminist and post-colonial theories, Code analyzes extended examples from developmental psychology, and from two "natural" institutions of knowledge production—medicine and law.

(Studies in Feminist Philosophy)

2006 paper $35.00 cloth $74 00

Philosophical Papers

Volume One

PETER UNGER

These two volumes of Unger's *Philosophical Papers* include articles spanning more than 35 years. The first volume collects work in epistemology and ethics, among other topics, while the second volume focuses on metaphysics.

2006 $55.00

Philosophical Papers

Volume Two

PETER UNGER

2006 $55.00

Oxford Studies in Epistemology

Volume 1

Edited by TAMAR SZABO GENDLER
and JOHN HAWTHORNE

Oxford Studies in Epistemology is a major new biennial volume offering a regular snapshot of state-of-the-art work in this important field.

(Oxford Studies in Epistemology)

2006 paper $29 95 cloth $74.00

Metaphysics, Mathematics, and Meaning

Philosophical Papers, Volume I

NATHAN SALMON

Metaphysics, Mathematics, and Meaning brings together Nathan Salmon's influential papers on topics in the metaphysics of existence, non-existence, and fiction; modality and its logic; strict identity, including personal identity; numbers and numerical quantifiers; the philosophical significance of Gödel's Incompleteness theorems; and semantic content and designation.

2006 paper $35.00 cloth $99 00

PRAGMATISM, OLD AND NEW

Selected Writings

Edited by Susan Haack, with Robert Lane

707 pp (Photos) • ISBN 1-59102-359-9 • HC • $32

This wide-ranging anthology tells the story of the evolution of pragmatism. Opening with a history of pragmatism from its inception to the present day, and closing with Haack's famous "interview" with Peirce and Rorty, the book presents a broad and diverse selection of pragmatist writings—classical and contemporary, reformist and revolutionary—on logic, metaphysics, theory of inquiry, philosophy of mind, philosophy of religion, aesthetics, philosophy of education, and moral, social, and political philosophy.

THE IMPROBABILITY OF GOD

Edited by Michael Martin and Ricki Monnier

380 pp • ISBN 1-59102-381-5 • HC • $32

This unique anthology collects most of the important arguments for the improbability of God that have been published since the mid-1900s. The editors make each argument clear and accessible by providing a helpful summary.

METAPHYSICS

The Key Issues from a Realistic Perspective

Nicholas Rescher

352 pp • ISBN 1-59102-372-6 • HC • $35

Metaphysics displays a clarity of exposition that makes it suitable for use as a text or supplementary reader in upper-class undergraduate and graduate philosophy courses.

THE ONTOLOGY OF TIME

L. Nathan Oaklander

366 pp • ISBN 1-59102-197-9 • PB • $30

"Oaklander's detailed responses to his many allies and opponents down the years also make his book an indispensable guide to the recent history of work in the philosophy of time." **—D. H. Mellor, Emeritus Professor of Philosophy Cambridge University**

KARL JASPERS'S PHILOSOPHY

Expositions and Interpretations

Edited by Kurt Salamun and Gregory J. Walters

275 pp • ISBN 1-59102-375-0 • HC • $50

The expository and interpretive essays in this volume, each uniquely commissioned, encompass the extraordinary richness of Jaspers's original philosophical thinking.

MARX, REASON, AND THE ART OF FREEDOM

Second Edition

Kevin M. Brien

Foreword by Marx W. Wartofsky

312 pp • ISBN 1-59102-366-1 • PB • $23

This rigorously argued and deeply thoughtful analysis reveals the continuing relevance and promise of Marx's thought in the 21st century.

Prometheus Books

Toll-Free: 800-421-0351 or Fax: 716-691-0137 www.prometheusbooks.com

Calcutta University Library
Datta
30.8.06
ALIPORE
CAMPUS

THE JOURNAL OF PHILOSOPHY

VOLUME CIII, NUMBER 2
FEBRUARY 2006

page
55 *Where Is the Theory in Our "Theories" of Causality?* Nancy Cartwright
67 *Rationality and the Ethics of Logic* Robert Hanna

COMMENTS AND CRITICISM

101 *Bays, Steiner, and Wittgenstein's "Notorious" Paragraph about the Gödel Theorem* Juliet Floyd and Hilary Putnam

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Christopher Peacocke, Carol Rovane, Achille C. Varzi. *Consulting Editors*: David Albert, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2006

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.
POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2006 by the Journal of Philosophy, Inc. ISSN 0022-362X

Calcutta University Library
ALIPORE CAMPUS

THE JOURNAL OF PHILOSOPHY
VOLUME CIII, NO. 2, FEBRUARY 2006

WHERE IS THE THEORY IN OUR "THEORIES" OF CAUSALITY?*

Causality is a hot topic today both in philosophy and in economics; there are approaching a dozen different theories of causality on offer. Few are reductionist—they do not embrace the Hume program of replacing causation by something weaker. But all claim to be "informative" theories, to tell us what are the central "characterizing" features of causation. Here is a list of just some of these theories and some of their major proponents. (I focus on theories of generic-level causation, or of "causal law." I choose this particular list because these are all theories that are closely related to practice and that I have studied fairly closely.) It certainly does not include all the accounts available.

"Theories" of Causality[1]

- The probabilistic theory of causality (Patrick Suppes) and its descendants
    - Bayes-nets theories (Wolfgang Spohn, Judea Pearl, Clark Glymour)
    - Granger causality (economist Clive Granger)

---

*This is a paper prepared for the 2005 annual conference of the British Society for the Philosophy of Science. This paper was also presented as the Nagel Lecture at Columbia University on September 22, 2005. I would like to thank members of both audiences for helpful discussion; also the British Academy for support.

[1] For more on these see: Patrick Suppes, *A Probabilistic Theory of Causality* (Amsterdam: North-Holland, 1970); Wolfgang Spohn, "Bayesian Nets Are All There Is to Causality" in Dominico Costantini, Maria Carla Galavotti, and Suppes, eds., *Stochastic Causality* (Stanford: CSLI, 2001), pp. 157–72; Judea Pearl, *Causality: Models, Reasoning, and Inference* (New York: Cambridge, 2000); Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, Prediction, and Search* (New York: Springer, 1993); James Woodward,

0022-362X/06/0302/55–66

- Modularity accounts (Pearl, James Woodward, economist Stephen LeRoy)
- Manipulation accounts (Peter Menzies, Huw Price)
- Invariance accounts (Woodward, economist/philosopher Kevin Hoover, economist David Hendry)
- Natural experiments (Herbert Simon, economist James Hamilton)
- Causal process theories (Wesley Salmon, Philip Dowe)
- The efficacy account (Hoover)
- Counterfactual accounts (David Lewis, Hendry, social scientists Paul Holland and Donald Rubin)

What is my worry about the theory in these theories of causality? Rather than explaining the worry directly as I now see it, I shall instead describe a web of thought that brought me to it. In the end I do not think one needs to share many of the views I will describe in order to share my concerns. But I think that working through them with me will make the problem more vivid and provide it with a kind of texture and depth that might otherwise be missing.

I. ON "THEORIES" OF CAUSALITY

I.1. Causation: One Word, Many Things. Generally when I open my eyes and look around me, I see a dappled world, plotted and pieced, not one homogenous sweeping plain. This tendency has been reinforced by my studies of causality over the last few years, which lead me to the conclusion that causation is not one monolithic concept; nor is there one single thing—the "causal relation"—that underpins our correct uses of that concept. There are a variety of different kinds of relations picked out by the abstract term "causes" and a variety of different—

Making Things Happen: A Theory of Causal Explanation (New York: Oxford, 2003); Stephen LeRoy, "Causality in Economics," *Causality: Metaphysics and Methods Technical Report*, CTR 20/04 (London: CPNSS, London School of Economics, 2004); Kevin Hoover, *Causality in Macroeconomics* (New York: Cambridge, 2001); Herbert Simon, "Causal Ordering and Identifiability" (1953), in *Models of Man* (New York: Wiley, 1957), pp. 10–35; James Hamilton, "Measuring the Liquidity Effect," *American Economic Review*, LXXI (1997): 80–97; Nancy Cartwright, *Nature's Capacities and Their Measurement* (New York: Oxford, 1989); Wesley Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton: University Press, 1984); Philip Dowe, *Physical Causation* (New York: Cambridge, 2000); David Lewis, "Causation," this JOURNAL, LXX, 17 (October 11, 1973): 556–67; David Hendry, "Causality and Exogeneity in Non-stationary Economic Time Series," *Causality: Metaphysics and Methods Technical Report*, CTR 18/04 (London: CPNSS, London School of Economics, 2004); Huw Price, "Agency and Probabilistic Causality," *British Journal for the Philosophy of Science*, XLII (1991): 157–76; Peter Menzies and Price, "Causation as a Secondary Quality," *British Journal for the Philosophy of Science*, XLIV (1993): 187–203; Clive Granger, "Testing for Causality: A Personal Viewpoint," *Journal of Economic Dynamics and Control*, II (1980): 329–52; Paul Holland and Donald Rubin, "Causal Inference in Retrospective Studies," *Evaluation Review*, XII (1988): 203–31.

correct—uses of the term for a variety of different purposes, with little of substantial content in common.

The variety of theories of causal law on offer provides one of the major reasons in favor of this plurality view. Each seems to be a good treatment of the paradigms that are used to illustrate it, but each has counterexamples and problems. Generally the counterexamples are provided by paradigms from some one or another of the other theories of causality. Usually in these cases we can see roughly why the one treatment is good for the one kind of example but not for the other, and the reverse. Sometimes we do better: we can give an explicit characterization of the kind of system of causal laws that the treatment presupposes and in the nicest cases we can prove a kind of representation theorem. (This is what I think we should always be aiming for.)

Bayes-nets are a good case. A Bayes-net account of what causality is, of the kind Wolfgang Spohn offers (*op. cit.*), is provably correct for a system of causal laws over a causally sufficient set of quantities[2] for which the three Bayes-nets postulates (causal Markov, minimality, and faithfulness) hold plus the assumption that if any two factors are probabilistically dependent then either one of them causes the other or they have a common cause. Similarly, I prove a representation theorem that shows that for systems of laws that satisfy the axioms of (what I call) a "linear deterministic system," James Woodward's (level-) invariance account of causation always yields correct results.[3]

If this account is correct then we have a variety of different kinds of causal systems and what we indicate by saying that one quantity causes another varies according to what characterizes the different systems. This makes a problem for method.

I.2. Causality: Metaphysics and Methods. Metaphysics and method should march hand-in-hand. If we have an account of what causation is then we ought to be able to explain why our methods for licensing causal claims are good methods for finding just the thing we say causality is. If our account of causation does not mesh with our best methods for finding out about causes, something has gone wrong on one side or the other.

---

[2] Roughly, this is a set of quantities such that "all" common causes of quantities in the set are themselves in the set. The scare quotes around "all" indicate that this needs a far more careful statement since all causes of common causes are themselves common causes and we may not want a condition that strong.

[3] See the theorem in section II.6 in Cartwright, *Hunting Causes and Using Them* (New York: Cambridge, forthcoming), and for this interpretation of the theorem, see Cartwright, "In Praise of the Representation Theorems," presented at the *International Lauener Symposium on Analytic Philosophy, in Honour of Patrick Suppes* (Berne, Switzerland, forthcoming).

We are used to thinking of causation as one thing, with many different methods to learn about it. Life is more complicated if we accept that there are different kinds of causation with different characterizing features. We must then face up to the question, which methods are appropriate for which kinds of causation? In fact for many of our philosophic theories of causality the question is not so difficult to answer since the theories themselves are almost read off from one method or another. Consider Patrick Suppes's probabilistic theory of causality (*op. cit.*), which forms the basis as well for Bayes-nets theories (my version of Suppes's theory).

> *Probabilistic "Theory" of Causation*
> THEORY: C causes E in K_i *iff* $P(E|C\&K_i) > P(E|\neg C\&K_i)$
> (K_i is a state description over a "complete"[4] set of confounding factors)
> METHOD: "C causes E" is licensed if C and E are probabilistically dependent once we have stratified on a "complete" set of confounding factors.

For varieties of causation where methods are not so immediately apparent, clearly it is a pressing matter to find appropriate methods and to show that they are appropriate. This was a project that a number of us worked on for several years, sponsored by the Arts and Humanities Research Board (AHRB).[5]

I.3. Representation: Handsome Is as Handsome Does. I argue for (borrowing an expression of Maria Carla Galavotti[6]) plurality in causality. But I still look for something in common. Is there not something in virtue of which it is correct to call all these different kinds of relations "causal"? Here is a proposal.[7] Rather than looking for one special relation in the world that legitimates representing them as causal, look instead for some unified features of the representations themselves. Clearly under this proposal we would not want to be thinking in terms of some kind of "Fido"-Fido or correspondence theory of representation.

---

[4] The scare quotes are around "complete" because it is a difficult notion to define. In my opinion it can only be defined relative to a particular choice of causal structure. That is why it has been so difficult to get an exact formulation for the general case. (See, for instance, all the difficulties laid out in Cartwright, *Nature's Capacities and Their Measurement*).

[5] AHRB-funded *Causality: Metaphysics and Methods*, Centre for Philosophy of Natural and Social Science, London School of Economics, 2002–2004. For more information see http://www.lse.ac.uk/collections/CPNSS/projects/ConcludedProjects/causality-MetaphysicsAndMethods.htm.

[6] Galavotti, "Plurality in Causality," presented at the *Seventh Meeting of the Pittsburgh-Konstanz Colloquium in the Philosophy of Science*, Konstanz, Germany (May 2005).

[7] This proposal is also discussed in section I.2 of Cartwright, *Hunting Causes and Using Them.*

So I turn instead to a theory of scientific representation championed by Mauricio Suarez[8] and Chris Swoyer.[9] A scientific representation should allow us to make inferences about the system represented. Whether or not this is a good starting point for thinking of scientific representation in general, it might be a help in thinking about what makes it right to call a variety of different kinds of relations all "causal." The reason for being optimistic about this suggestion is that there is a certain kind of inference that has always been stressed as central to the notion of causation, to which I turn next.

I.4. Causal Laws and Effective Strategies. The idea that causes allow us to affect the world has always been part of our thinking about the notion of cause, it is at the core of our current manipulation theories as well as some invariance theories, it is probably the dominant idea about causation in economics right now, and it was right at the fore when talk of causation first reappeared defying Positivist prohibition, as witnessed by the title, "Causal Laws and Effective Strategies," of my first paper on causality.[10] The idea suggested by this in combination with the inference-based approach to representation is that representing a set of relations as causal should allow us to make some kinds of inferences that allow us to use causes as strategies for producing effects. Is something like this possible?

Let me start with some theories of causality where such inferences hold center place. I shall illustrate with two theories coming from economists rather than any of the manipulation theories of philosophers for reasons that should be apparent later. Kevin Hoover, in *Causality in Macroeconomics* (*op. cit.*), distinguishes between quantities we control directly and those we do not. His aim is to characterize the causal relations among the latter using the concept of "direct control."[11] He then takes the following as the central characterizing feature of causality:

> *Hoover*:
> x causes y iff anything we can do to fix x affects y but not the reverse.

---

[8] Suarez, "Theories, Models and Representations," in Lorenzo Magnani, Nancy Nersessian, and Paul Thagard, eds., *Model Based Reasoning in Scientific Discovery* (New York: Kluwer, 1999), pp. 75–83.

[9] Swoyer, "Structural Representation and Surrogative Reasoning," *Synthese*, LXXXVII (1991): 449–508.

[10] Cartwright, "Causal Laws and Effective Strategies," *Noûs*, XIII (1979): 419–37.

[11] For more about Hoover's proposal, see section III.4 in Cartwright, *Hunting Causes and Using Them*.

If that is what causality is, then the inference from "x causes y" to "We can influence y by fixing x" is built right into the metaphysics.

Similarly for econometrician David Hendry.[12]

Hendry:

- Causes are superexogeneous.
- Superexogeneity = weak exogeneity + "invariance"

WEAK EXOGENEITY:
Given $P(y\&x, \beta, \gamma) = P(y|x,\beta)P(x,\gamma)$,
x is weakly exogeneous to a vector of outcome variables y if the vector of parameters γ of the marginal distribution of x have no cross-restraints with the vector of parameters β of the conditional distribution of y on x.[13]

SUPEREXOGENEITY:
In addition the conditional distribution is invariant to interventions that change (parameters of) the marginal distribution.

Suppose then that x is weakly exogenous with respect to y and we think of changing the distribution of x in order to affect the distribution of y. It may seem that we can predict the outcome from the formula for the conditional distribution. But that is not so. Changing changes the joint probability distribution and there is nothing that ensures that the new conditional distribution will still be the same. In the original distribution they may have no dependencies but that does not show what happens if the distribution is changed. So Hendry adds the constraint:[14] x causes y only if the parameters, β, of the conditional distribution, $P(y|x)$, stay fixed as we vary the parameters, γ, of the distribution of x. Again an inference about use follows immediately from the very notion of causation.

Can we do this in general? Recall, that was my hope: representing a relation as causal allows some kinds of inferences about use. Unfortunately, this hope is not supported by the other theories of causality on offer. To see this vividly consider Woodward's manipulation/invariance theory of causality, a case where we might expect the connection to be immediate just as in the Hoover and Hendry theories.

---

[12] See Robert Engle, Hendry, and Jean-François Richard, "Exogeneity," *Econometrica*, LI (1983): 277–304, and Hendry, "Causality and Exogeneity in Non-stationary Economic Time Series," plus the discussions of Hendry in section III.6 in Cartwright, *Hunting Causes and Using Them.*

[13] This is only a special case of the definition, for illustration. Weak exogeneity implies that the marginal distribution can be ignored in estimating the conditional distribution.

[14] Note that Hendry's position differs from Hoover's in that he provides a necessary condition for the causal relation rather than an outright definition.

I.5. Hunting Causes versus Using Them. It is often supposed[15] that the "gold standard" for hunting causes—that is, for ascertaining causal relations—is the *ideal Galilean experiment.* In this kind of experiment we vary the cause holding fixed "all" other sources of variation of the putative effect and look to see if the effect varies in train. This is just what Woodward uses to characterize causation.[16] He considers situations in which the putative effect, *y*, is a function of the putative cause, *x*; so they co-vary. If *x* is to be correctly said to cause *y* then he demands the very strong condition that *x* be able to vary independently of all other sources of variation in *y* as we should like to see in a Galilean experiment. This condition is called *modularity.* He also requires that the functional relation between *x* and *y* stay the same across variations in *x* that keep fixed the other causes of *y*. This ensures that *y* will co-vary with *x* in a Galilean experiment, as required if *x* is to be judged a cause of *y* from the experiment (but it is stronger than required to ensure the covariation). This condition is called *invariance.* So...

> *Woodward:*
> Given $y = f(x,\ldots)$, x causes y *iff*
> (1) modularity: x can be varied independently of all "other" sources of variation in y
> (2) invariance: $y = f(x,\ldots)$ continues to hold as x is varied holding fixed other sources of variation in y

This might reasonably be called a *manipulation account* but it is very different from the two manipulation accounts from economics that I have just reviewed. Woodward's theory of causality licenses counterfactuals about what would happen if the cause were to vary in a very special way, a way that is not what we would normally be envisaging for either policy or technology. The variation is just the kind we need if we want to test a causal claim—to hunt for causes—but not the kind we expect to implement when we try to use them. The same is true for many of the other accounts. Recall that many are almost read off some one or another method for establishing causal claims. Correlatively, like Woodward's account they give little guidance about what we can do with a causal claim once we know we are entitled to it.

---

[15] Wrongly I think. See sections I.3, II.3–6, and III.6 in Cartwright, *Hunting Causes and Using Them.*

[16] This is my interpretative summary of what Woodward's account amounts to in brief and not literally how he defines things. See Woodward, *Making Things Happen,* and my discussion of his invariance views in sections II.2–6 in Cartwright, *Hunting Causes and Using Them.*

This observation is true almost universally across our list of causal theories. They are in essence either "language-entry rules"—they tell us when it is correct to label relations "causal" (and generally only one method for so doing)—or they are "language-exit rules" that tell us what inferences we can draw from that label (and again, usually only one particular kind of inference). This raises the twin concerns:

- What is the use of causes characterized from a hunting methodology?
- How do we hunt causes when they are characterized in terms of their use?

I.6. Causality: Metaphysics, Methods, and Use. I have briefly sketched the first of my twin problems in referring to Woodward's manipulation/invariance account. The same holds for David Lewis-style counterfactual accounts. They provide an elaborate procedure for deciding when we can attach the label "cause." But then what? There is nothing more in the account that allows us to move anywhere from that, nothing that licenses any inferences for use.

The second concern can be readily seen in Milton Friedman's famous arguments in favor of hugely idealized models in economics.[17] Models, he urges, should be judged by the correctness of their results in the domain under investigation, not by the correctness of their assumptions. So like the economists' accounts of causality I discussed above, for Friedman models are geared to use. But this very fact restricts their usefulness. The most useful models are ones we can first legitimate by some independent means and then use to make predictions about new matters. On Friedman's account all there is to the acceptability of the model is that it gives correct predictions about the matters we are interested in—and that is just what we cannot know about the model beforehand if our point in using it is to gain confidence in those predictions. Similarly for Hoover's account of causality. What we should like is to have an independent reason for thinking that *x* causes *y*, then use that to infer that manipulating *x* will affect *y*. But for Hoover that is just what it is for *x* to cause *y* so we cannot establish the causal relation first and then use it later for policy predictions.

What we see from the twin concerns above is that in all the "theories" of causation listed, the metaphysics is too thin. It is geared either to hunting causes or to using them but not to both. My own earlier project, "Causality: Metaphysics and Methods," missed this point. Yes, the metaphysics of causality must be tied to methods on the one hand, but it must equally be tied to use on the other. As I urge in the introduction to my latest book,[18]

---

[17] Friedman, "The Methodology of Positive Economics," in *Essays in Positive Economics* (Chicago: University Press, 1953), pp. 3–43.

> Our philosophical treatment of causation must make clear why the methods we use for testing causal claims provide good warrant for the uses to which we put those claims.

I should stress that this same requirement holds equally for a universalist and for a pluralist view of causality. Our notions of causation, or our characterizations of specific kinds of causal systems, must be rich enough to allow us both to ascribe the label "causal" and then to make some use of it once we have done so.

I.7. Where Is the Theory in Our "Theories" of Causality? Considering Hendry in this light raises yet another issue. Weak exogeneity is a concept having to do with statistical inference—estimating what a distribution is. This has little apparent connection with causality. But Hendry's attention to the conditional distribution of y on x—$P(y|x)$—in his account of "x causes y" is reminiscent of the probabilistic theory of causality. Then he adds on top an invariance condition. It is not that we do not have available any methods for testing for the kind of invariance that Hendry demands. Econometrics has good methods for looking for "structural breaks"—places where a distribution changes. The idea then is that we look to see if structural breaks in the distribution we will use for prediction are associated with changes in the parameter we propose to manipulate for policy.

What is disappointing from the point of view of the metaphysics of causality is that the invariance condition is, as I said, added on. And it is what does the work of licensing predictions for use, not any special features of the thing that is itself invariant. Any kind of formula that allows us to predict what we want to know and that holds true across the variations we envisage making will do; it need not have any of the other kinds of characteristics that are in the cluster we usually associate with causality—increase in probability of the effect on the cause, space-time connection, asymmetry of the cause-effect relation, existence of a mechanism, and so on.[19]

---

[18] See Cartwright, *Hunting Causes and Using Them.*

[19] We can see a good instance of this in my discussion of Hoover's theory of causality in section III.4 of Cartwright, *Hunting Causes and Using Them.* There I argue that Hoover's "strategy" relations are distinct from "production" relations, which have many of the other features generally associated with causality. On the other hand, the kinds of production relations described there can have just the drawback I am worrying about here—nothing seems to follow from them about strategy, except the truism that if the equations continue to be accurate across the changes we will make, they can be used for prediction of what would happen were we to make those changes.

This is the point that Sandra Mitchell makes about the notion of "scientific law."[20] For use in prediction we do not need a claim that holds universally; nor, it seems, do we need anything that satisfies any of our usual accounts of what it is to be a "causal" law. We only need a claim that holds across the range of situations in which we will use it for prediction. Sometimes the same point seems to be urged by Woodward too. He seems to argue that invariance—and just invariance—across the required, possibly highly limited, range of envisaged uses is what matters. We might call that "causation" then. But he does not embrace this: he insists on the Galilean manipulation account of causation that I described, which is not geared to practical use.

Bayes-nets methods make some advance here. I think for instance of the manipulation theorem of Peter Spirtes, Clark Glymour, and Richard Scheines, that computes facts about the joint probability of sets of variables after manipulation from facts about their distribution before plus facts recorded in the causal graph, or of Judea Pearl's thick book on counterfactuals (*op. cit.*). I worry though that both cases are still too close to the strategy I ascribed to Hendry—the "add-on" strategy.

Bayes-nets characterize causality in terms of the three axioms I cited, the chief of which is the causal Markov condition. But how important is the causal Markov condition to our ability to predict facts about the probability distribution after manipulation from knowledge of the causal relations plus the probability distribution before? I do not think it can be too important.

Consider for example what I call "epistemically convenient systems" with probability measures.[21] For these we can define a very close analogue of the Spirtes-Glymour-Scheines concept of manipulation, including, as they do, the assumption that the causal relations are invariant under manipulation. Systems like this need not satisfy the causal Markov condition. Yet we can derive results about what happens to the joint distributions under manipulation in these kinds of systems just as we can when the causal Markov condition is satisfied.[22] In both cases most of the work in deriving the relevant results is done not by the axioms of the system but by the very restricted notion of

---

[20] Mitchell, *Biological Complexity and Integrative Pluralism* (New York: Cambridge, 2003).

[21] See section II.6 of Cartwright, *Hunting Causes and Using Them.*

[22] What the causal Markov condition does is to give the joint probabilities both before and after a particularly nice form. Roughly, conditioning on the direct causes of *x*, *y* becomes independent of *x* if *x* does not cause *y*, which means that we compute the probability of *y* without regard to *x*.

"manipulation" that is used. The manipulations involved in these theorems are like Woodward's (or those miracle manipulations of Lewis)—the kinds of manipulations we should like to make in a Galilean experiment, not the kind we envisage for policy, where many things may be expected to change at once, including some of the causal principles by which the system operates.[23] The situation is similar for Pearl on counterfactuals: most of the work in legitimating policy predictions comes from the semantics he offers for counterfactuals, not from the special modularity conception of causality he urges (though the two are nicely meshed).[24]

More generally, all the cases that I know where we have counterfactual inferences or manipulation theorems that can be proven from a particular kind of system of causal principles work like this: first, a set of relations are postulated that satisfy a particular set of constraints taken to be sufficient to characterize them as "causal relations." Second, manipulations are taken to create a new, closely related system. The new systems are defined by essentially two clauses. The first describes what is to change, for example, the value of certain variables, maybe representing quantities all of which can vary independently of each other (that is, they are "variation-free"), or the relation by which the "manipulated" quantity is produced. The second states a set of things that do not change, like values of other particular variables and all the other specified relations except that by which the "manipulated" quantity is produced. Then the counterfactual consequence is calculated in the new system.

What is striking about this is that the constraints that justify labeling the system of relations as "causal" seem to have no role to play. What matters is that there is a system of relations, say equations, that allow us to solve for a given result. Then we are given a recipe for making very specific changes and the recipe produces a new system that also allows us to solve for the result in question. The solvability of the two systems plus the rule for getting from one to the other is what does the job, not any of the constraints that make it a "causal" system in the first place (beyond those that ensure solubility). So what in the end does causation have to do with policy counterfactuals, manipulation, or effective strategies?

---

[23] Cf. the famous Lucas critique of macroeconomics, which depends on the assumption that manipulating policy variables produces changes in the causal principles governing the system. See Robert Lucas, "Econometric Policy Evaluation: A Critique" (1976), in *Studies in Business Cycle Theory* (Cambridge: MIT, 1983), pp. 104–30.

[24] And, like Woodward and Lewis, that semantics is geared to hunting causes, not using them.

I have been hoping for something more, something that shows why our methods should warrant causal claims with practical strategy implications—and something of substance that does so. It is not enough for a metaphysics of causality to adopt just the "add-on" strategy, to be simply a union of a language entry rule and a language exit rule. What we should be looking for is a *theory* of causality, in much the same way as we have a theory of the electron. The account itself should combine with other chunks of knowledge to imply both the language entry and the language exit rules.

II. CONCLUSION

When it comes to theories of causal law, I have earlier urged that we have on offer not a plurality of competing theories but rather a host of theories each appropriate to one of a large plurality of different kinds of (systems of) causal laws. I now worry that this was too generous. Rather than a plurality of theories, we do not have any theories at all.

I close with a hypothesis to think about. Bertrand Russell maintained that there are no causal claims in physics.[25] Many of us have urged the contrary. Physics is rife with causal claims, for instance Einstein and Infeld: "Forces cause motions: Gravitational forces along the line of action; electromagnetic, parallel to the line of action." If so then what we have in each of the specific domains of physics are theories about that domain, theories that contain causal claims. "Massive bodies attract each other." Obviously the same could be said across the sciences. Perhaps we should find our lead here. Maybe our whole enterprise is misplaced. Perhaps we should be seeking not *theories of causality* but rather *causal theories of the world.*

NANCY CARTWRIGHT

London School of Economics
University of California, San Diego

---

[25] Russell, "On the Notion of Cause," *Proceedings of the Aristotelian Society*, XIII (1913): 1–26.

RATIONALITY AND THE ETHICS OF LOGIC*

> Logic is a science of reason...a science a priori of the necessary laws of thought, not in regard to particular objects, however, but to all objects in general;—hence a science of the correct use of the understanding and of reason in general, not subjectively, however, i.e., not according to empirical (psychological) principles for how the understanding does think, but objectively, i.e., according to principles a priori for how it ought to think.
>
> —Immanuel Kant[1]

> [T]he word Logic in its primal sense means the Science of the Laws of Thought as expressed. Considered in this sense, Logic is conversant about all thought which admits of expression; whether that expression be effected by the signs of common language or by the symbolic language of the mathematician.
>
> —George Boole[2]

> The logical notions are embedded in our deepest nature; in the very form of our language and thought, which is presumably why we can understand some kinds of logical systems quite readily, whereas others are inaccessible to us without considerable effort...if at all.
>
> —Noam Chomsky[3]

> If sheer logic is not conclusive, what is?
>
> —W.V.O. Quine[4]

> Which logic do we use to assess the consequences of different logics?... Regress threatens. Is the super-logic...a priori, or incorrigible?
>
> —Stewart Shapiro[5]

---

* I am very grateful to Michael Potter for many conversations, in many pubs, on and around the topics of this essay.

[1] Kant, *The Jäsche Logic, Lectures on Logic*, J.M. Young, trans. (New York: Cambridge, 1992), pp. 521–640, at p. 531.

[2] Boole, "Logic and Reasoning," in *Studies in Logic and Probability* (London: Watts, 1952), pp. 211–29, at p. 212.

[3] Chomsky, *Language and Problems of Knowledge* (New York: Columbia, 1988), p. 99.

[4] Quine, *Philosophy of Logic* (Cambridge: Harvard, 1986, 2nd ed.), p. 81.

[5] Shapiro, "The Status of Logic," in Paul Boghossian and Christopher Peacocke, eds., *New Essays on the A Priori* (New York: Oxford, 2000), pp. 333–66, at p. 338.

0022-362X/06/0302/67–100

> When studying reasons we study normative aspects of the world. When discussing rationality we discuss our perceptions of, and responses to, reasons. Our ability to reason is central to our rationality in all its manifestations.
>
> —Joseph Raz[6]

> Logic and ethics are fundamentally the same, they are no more than duty to oneself.
>
> —Otto Weininger[7]

This essay is about human rationality, logic, and the connection between them. On my view, this connection is both constitutive and mutual. More precisely, I defend the broadly Kantian thesis[8] that logic is the result of the constructive operations of an innate cognitive capacity that is necessarily shared by all rational human animals, and governed by categorically normative principles. I call this thesis *Kantian constructivism in logical theory* (KCLT).[9] In section I, I sketch a picture of the nature of logic that connects it directly with human psychology. Here I propose that all rational humans innately possess a cognitive faculty that is preconfigured for representing logic by virtue of its containing a single universal "protologic" which is used for the construction of all logical systems, and which is distinct in structure from all classical or nonclassical logical systems. In section II, I argue that logic is intrinsically normative, that the intrinsic normativity of logic is generally consistent with logic's being the science of the necessary relation of consequence, and in particular that the intrinsic normativity of logic is consistent with the existence of the protologic. In section III, I argue that the protologic also provides a set of unconditional prescriptive laws, or "categorical imperatives," for human reasoning. Finally, by way of conclusion, I briefly argue that my account of the intrinsic categorical normativity of logic supplies a crisp and effective reply to an obvious worry: If rational human animals are inherently logical animals, then how could they ever make logical mistakes?

---

[6] Raz, *Engaging Reason* (New York: Oxford, 1999), p. 75.

[7] Weininger, *Sex and Character* (London: Heinemann, 1906), p. 159.

[8] By this I mean only that that my theory of rationality and logic shares some basic themes and theses with certain parts of Kant's Critical philosophy. But of course my theory is intended to be independently philosophically justified and motivated; and also it is not in any way intended to be an interpretation of or commentary on Kant's logical writings.

[9] See John Rawls, "Kantian Constructivism in Moral Theory," this JOURNAL, LXXVII, 9 (September 1980): 515–72, and sections II-III below.

Quite apart from the Kantian affinities, all of this is well supported by contemporary cognitive science and logical theory. But if I am correct, then there are also some fairly radical foundational implications for cognitive science and logical theory alike.

I. THE RATIONAL, THE LOGICAL, AND THE PSYCHOLOGICAL

As I said in the Introduction, the central claim of this essay is the thesis of KCLT: logic is the result of the constructive operations of an innate cognitive capacity that is necessarily shared by all rational human animals, and governed by categorically normative principles. KCLT connects the logical and the psychological intrinsically and reciprocally. But the dual idea that logic is intrinsically psychological and that human psychology is intrinsically logical has a long and troubled history in the philosophical, logical, and psychological traditions alike. From Antoine Arnauld's and Pierre Nicole's *Art of Thinking* (1662), through Immanuel Kant's *Jäsche Logic* (1800), J.S. Mill's *System of Logic* (1843), and George Boole's *Investigation of the Laws of Thought* (1854), right up to the appearance of Gottlob Frege's revolutionary *Begriffsschrift* (1879), logic and psychology seemed to be, if not precisely the same subject, then at least theoretically married to one another. But the much-celebrated attack on "logical psychologism"—the explanatory reduction of logic to empirical psychology—at the end of the nineteenth century brought about a nasty divorce. According to the leaders of the attack, Frege and Edmund Husserl, this parting of the ways was a simple matter of irreconcilable differences: the principles or laws of logic are absolutely necessary while the laws of empirical psychology are only contingent generalizations; logic is true while empirical psychology deals only with human belief; logic is a fully formal or "topic-neutral" science while empirical psychology focuses only on the species-specific or individual contents of mental states; logical knowledge is a priori or independent of all sense experience while empirical psychological knowledge is a posteriori or dependent upon experience; and so on. Thereafter "pure logic," pursued in armchairs by philosophers and philosophically-minded mathematicians, went one way; and "experimental psychology," pursued in laboratories by men in white coats, went diametrically another. To make things worse, as Elliott Sober aptly observes, "while the psychologists were leaving, the philosophers were slamming the door behind them."[10]

---

[10] Sober, "Psychologism," *Journal for the Theory of Social Behavior*, VIII (1978): 165–92, at p. 165.

In my opinion the view that logic and psychology are fundamentally at odds with one another could not be more mistaken. On the contrary, if I am correct there is a necessary link between logic and psychology, despite the fact that logical psychologism is false. This brings me back to the first part of KCLT: *that logic is cognitively constructed by rational human animals,* in the sense that all rational humans innately possess a cognitive faculty that is preconfigured for representing logic, by virtue of its containing a single universal *protologic* which is used for the construction of all logical systems, and which is distinct in structure from all classical or nonclassical logical systems. I call this subclaim *the logic faculty thesis.* The logic faculty thesis draws explicitly but not uncritically on some ideas of Kant, Boole, W.V. Quine, Noam Chomsky, and Stewart Shapiro.

Obviously the fundamental notion lying behind the logic faculty thesis is that of a rational animal. For my purposes animals are *sentient living organisms* and for simplicity's sake I shall assume unless otherwise specified that all animals are *sound,* that is, intact and mature. Even so, only some animals in this sense are rational. On my view, rational animals are rule-following,[11] intentional (that is, possessing capacities for object-directed cognition and purposive action), volitional (possessing a capacity for willing),[12] self-evaluating, self-justifying, self-legislating, reasons-giving, reasons-sensitive, and reflectively self-conscious—or for short, "normative-reflective"[13]—animals, whose inner and outer lives alike are sharply constrained by their possession of concepts expressing strict modality. Modality in the philosophical sense comprises the concepts of necessity, possibility, and contingency. Strict modality, in turn, includes the concepts of logical necessity (truth in all logically possible worlds), epistemic necessity (certainty or indubitability), and deontic necessity (unconditional obligation or "the ought"). So to put the first part of KCLT yet another way, logic is cognitively constructed by all and only those human normative-reflective animals who are also in possession of concepts expressing strict modality.

This approach to rational animals substantively invokes the concept of rationality. An unfortunate but pervasive feature of the

---

[11] See Jonathan Bennett, *Rationality* (Indianapolis: Hackett, 1989).

[12] See Harry Frankfurt, "Freedom of the Will and the Concept of a Person," in *The Importance of What We Care About* (New York: Cambridge, 1988), pp. 11–25; and Brian O'Shaughnessy, *The Will: A Dual Aspect Theory,* 2 Volumes (New York: Cambridge, 1980).

[13] See Robert Audi, *The Architecture of Reason* (New York: Oxford, 2001), chapters 1–2 and 4–5; and Raz, chapters 1–4.

philosophy of rationality, however, is that it does not operate with either a univocal or generally-accepted sense of the term 'rationality'.[14] Reasonable people, including specialists on rationality, are both muddled and also in sharp disagreement about the very concept of rationality. So in order to avoid troublesome ambiguity and state my commitments explicitly I need to make some basic distinctions, and orient my view in relation to them.

The first basic distinction is between (a) the *mentalistic* sense of rationality, and (b) the *procedural* sense of rationality. In the mentalistic sense, rationality is a complex psychological capacity for logical inference and insight, and also for practical deliberation and decision-making. By contrast, in the procedural sense, rationality is a complex formal property of a certain class of mechanical, mathematical, computational, or logical processes, namely the property of being (i) well formed and (ii) either provable and recursive (Turing-computable), valid (truth-preserving), or sound (valid with true premises).[15] The crucial difference here is that rationality in the mentalistic sense is such that all of its manifestations are conscious, whereas some processes are rational in the procedural sense without being in any way conscious.

It is also quite useful to distinguish, within the mentalistic sense of rationality, between (a_1) the mentalistic rationality of *animals*, (a_2), the rationality of mental *episodes or acts*, and (a_3) the rationality of mental *states*. The important contrast here is that it is possible for something to be a rational animal by having an overall mental capacity for rationality, yet fail to be occurrently rational with respect to some of its mental episodes or mental states, as in the case of someone who completely loses his temper temporarily. Conversely, it is possible for an animal to be occurrently rational with respect to some of its mental episodes or states, but lack an overall mental capacity for rationality, as in the case of certain sorts of mental illness. This point in turn implies another useful distinction, again within the mentalistic sense of rationality, between (a_5) an animal's mental

---

[14] See, for example, Alfred R. Mele and Piers Rawling, eds., *The Oxford Handbook of Rationality* (New York: Oxford, 2003).

[15] See George Boolos and Richard Jeffrey, *Computability and Logic* (New York: Cambridge, 1996, 3rd ed.). Gödel's incompleteness theorems may be taken to show that validity and soundness outrun provability and computability. The precise character of logical validity will vary according to whether the logic is classical or nonclassical, and within nonclassical logics, according to whether the logic is either merely a conservative extension of classical logic or a deviant logic. What this means is that not all rational procedures are classically logically valid.

capacity for rationality, and (a_6) *occurrent* rationality with respect to the mental episodes or mental states of an animal. And finally, for completeness, we can also distinguish, within occurrent mentalistic rationality, between (a_7) the occurrent rationality of mental episode or state *types*, and (a_8) the occurrent rationality of mental episode or state *tokens*. Here the contrast is that it is possible for a certain mental episode or state type—say, righteous anger—to be rational when tokened in in some contexts, but fail to be rational when tokened in others.

The second basic distinction is between (c) the *meeting-the-minimal-standards* sense of rationality, and (d) the *meeting-the-maximal-or-ideal-standards* sense of rationality. In the meeting-the-minimal-standards sense, rationality means either possessing a psychological capacity for rationality or meeting the well-formedness conditions for being a rational procedure of the relevant sort. By contrast, in the meeting-the-maximal-or-ideal-standards sense, rationality means either perfectly using a psychological capacity or else perfectly satisfying the provability/computability-conditions, validity-conditions, or soundness-conditions of the relevant sort of rational procedure. The crucial difference here is that in the meeting-the-minimal-standards sense, irrationality means lacking the basic conditions necessary for rationality, hence *nonrationality*; whereas in the meeting-the-maximal-or-ideal-standards sense, irrationality merely means falling short of perfect rationality.

The third and last basic distinction is between (e) the *principled* sense of rationality, (f) the *holistic* sense of rationality, and (g) the *instrumental* sense of rationality. In the principled sense, rationality means the possession of a capacity for generating or recognizing necessary truths, a priori beliefs, strictly universal normative rules, nonconsequentialist moral obligations, and categorical "ought" claims.[16] Put in historical terms, this is the *Kantian* conception of rationality, acccording to which "reason is the faculty of a priori principles." By contrast, in the holistic sense, rationality means the possession of a capacity for systematically seeking coherence (or, to use a familiar term-of-art, "reflective equilibrium") across a network or web of beliefs, desires, emotions, intentions, and volitions.[17] In historical terms, this is the *Hegelian* conception of rationality, according to which "the truth is the whole." And finally in the instrumental sense, rationality means the possession of a capacity for generating or recognizing contingent truths, a posteriori beliefs, contextually normative rules, consequentialist obligations, and hypothetical "ought" claims.[18] Put historically, this is the *Humean* conception of rationality, according to which "reason is the slave of the passions."

---

[16] What I am calling "principled rationality" is closely connected with the traditional notion of "pure reason"; see, for example, Laurence Bonjour, *In Defense of Pure Reason* (New York: Cambridge, 1998).

[17] See Rawls, "The Independence of Moral Theory," *Proceedings and Addresses of the American Philosophical Association*, XLVIII (1974): 5–22; and Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971).

The crucial three-way difference here is that whereas in the principled sense, rationality means generating or recognizing rules that are absolute or *unconditional*, by contrast rationality in the holistic sense means generating or recognizing rules or laws that are merely thoroughly interdependent or *mutually conditioned* (hence none of those rules or laws can be more necessary or certain or binding than the modally or epistemically weakest proposition in the total holistic network of rules or laws), and by another contrast rationality in the instrumental sense means generating or recognizing rules that are merely empirically regular or *conditional* (hence none of those rules or laws can be fully necessary or certain or binding).

Unless otherwise noted, in what follows I will focus primarily on the mentalistic, meeting-the-minimal-standards, and principled senses of rationality. This is not to say that I reject or wish to depreciate in any way the procedural, meeting-the-maximal-or-ideal-standards, holistic, or instrumental senses of rationality. On the contrary, I am saying only that rationality in the senses I am primarily interested in *should not be confused with* other fundamentally different senses of rationality.

Now the class of normative-reflective animals in possession of concepts expressing strict modality would appear to be at least extensionally equivalent with the class of rational humans; and even if (as seems very likely) it is not intensionally equivalent for the simple reason that the cognitive capacities required for the possession of concepts expressing strict modality are multiply embodiable,[19] nevertheless those humans who are rational will constitute a central case or paradigm. I am assuming that it is a primitive fact, yielded directly by the reader's capacity for introspection, that there *are* some rational humans. So I am proposing to explain the nature of logic by taking human rationality seriously. More precisely, I am proposing to explain the nature of logic by taking rationality seriously, and to take rationality seriously by taking human rationality seriously. And what we reach at the end of this explanation is the thesis that something protological is built innately into human rationality itself.

---

[18] See Martin Hollis and Robert Sugden, "Rationality in Action," *Mind*, CII (1993): 1–35; Donald C. Hubin, "The Groundless Normativity of Instrumental Rationality," this JOURNAL, XCVIII, 9 (September 2001): 445–68; Robert Nozick, *The Nature of Rationality* (Cambridge: Harvard, 1993); and Herbert Simon, *Reason in Human Affairs* (Stanford: University Press, 1983).

[19] For me, something is "multiply embodiable" just in case it is a cognitive structure that can occur in two or more distinct biological individuals, natural kinds, or artificial living kinds. A good intuitive example of the latter would be the "Nexus VI replicants" in Ridley Scott's classic science fiction film *Blade Runner* (1982).

Logic, by which I mean *formal* logic, is the science of the necessary relation of consequence between the premises and the conclusion of valid arguments: there is no possible world such that all the premises are true and the conclusion false.[20] Since at least the time of C.I. Lewis's *Survey of Symbolic Logic*, it has been known that the "second-order" or "higher-order" logic of Russell and Whitehead's *Principia Mathematica* (so-called because it permits quantification into predicates and over properties, functions, and sets) can be modified in various ways.[21] This remains true, and true in spades, for classical or elementary logic, that is, bivalent truth-functional first-order quantified polyadic predicate calculus with identity. In her pathbreaking 1974 book *Deviant Logic*, Susan Haack helpfully collected these modifications together under the comprehensive heading of "alternative" or "non-classical" logics.[22] And she also very helpfully distinguished, within nonclassical logics, between (1) extensions of classical or elementary logic, and (2) deviant logics.

Briefly put, extensions of elementary logic introduce nontrivial changes (changes other than mere notational variation) that preserve all the logical constants, valid sentences, theorems, valid inferences, and laws of elementary logic. By contrast, deviants of elementary logic introduce nontrivial changes that do *not* preserve all the classical or elementary logical constants, valid sentences, theorems, valid inferences, and laws. Less briefly put, an extension of classical or elementary logic involves the addition, deletion, or redefinition of classical logical constants, interpretation rules, axioms, or inference rules such that all the tautologies, theorems, valid inferences, and laws of elementary logic still hold, along with some additional ones. For instance, the "calculus of strict implication" devised by Lewis,[23] and the different axiomatic modal logics developed by Lewis,[24] Saul Kripke,[25] and others, are extensions in this sense.

---

[20] Philosophers of logic often distinguish between *syntactic consequence* or provability, and *semantic consequence* or necessary truth-preservation. I accept that distinction, but for convenience will use 'consequence' simply to *mean* semantic consequence.

[21] See Lewis, "Alternative Systems of Logic," *Monist*, XLII (1932): 481–507; and Rudolf Carnap, *The Logical Syntax of Language*, Amethe Smeaton, trans. (London: Routledge and Kegan Paul, 1937).

[22] See Haack, *Deviant Logic* (New York: Cambridge, 1974), chapter 1; Haack, *Deviant Logic, Fuzzy Logic* (Chicago: University Press, 1996), chapters 1 and 9; and Graham Priest, *Introduction to Non-Classical Logic* (New York: Cambridge, 2001).

By contrast, a deviant logic involves the addition, deletion, or redefinition of classical logical operators, interpretation rules, axioms, or inference rules such that *not* all the tautologies, theorems, valid inferences, and laws of classical or elementary logic still hold. Deviant logics include intuitionist logics, which reject the classical or strong law of excluded middle;[26] relevance logics, which reject the classical conditional or classical validity;[27] three-valued, many-valued, truth-value-gapped, and fuzzy logics, which all reject the classical or strong law of bivalence;[28] "free" and "meinongian" logics, which reject unrestricted existential generalization;[29] paraconsistent logics, which allow contradictions to occur as theorems without entailing every sentence; and dialetheic logics, which are both paraconsistent *and* reject the classical or strong laws of noncontradiction and bivalence by allowing some truth-value "gluts" or true contradictions.[30] Extensions of elementary logic, while promoting some genuine changes to classical logic, are conservative. But a deviant logic is radical and *flouts* classical or elementary logic in one way or another.

This vividly raises the question: Given the various nonclassical logics and especially the deviant logics, what should we then say about the nature of logic?

Most philosophers of logic hold that there are three and only three mutually exclusive options for answering this question. The first option is to be a *diehard classicist* and insist that classical or elementary logic is the One True Logic and that the nonclassical logics are not really and truly logic. The second option is to be a *diehard nonclassicist* and insist that one or another of the nonclassical logics is the One True Logic and that all other logics including classical or elementary logic are not really and truly logic. And the third option is to be an *unconstrained pluralist* by (a) denying that there is any One True Logic, whether classical or nonclassical, and (b) asserting that the several different logics are all equally acceptable.[31]

---

[23] See Lewis, *Survey of Symbolic Logic* (Berkeley: California UP, 1918), chapter V.

[24] See Lewis and C.H. Langford, *Symbolic Logic* (New York: Dover, 1959, 2nd ed.).

[25] See Kripke, "Semantical Analysis of Modal Logic I, Normal Propositional Calculi," *Zeitschrift für mathematische Logik and Grundlagen der Mathematik*, IX (1963): 67–96; and Kripke, "Semantical Considerations on Modal Logic," *Acta Philosophica Fennica*, XVI (1963): 83–94.

[26] See L.E.J. Brouwer, "Historical Background, Principles, and Methods of Intuitionism," *South African Journal of Science*, XLIX (1952): 139–146; Haack, *Deviant Logic*, chapter 5; and Priest, *Introduction to Non-Classical Logic*, chapter 6.

[27] See Priest, chapters 9–10.

[28] See Haack, *Deviant Logic*, chapter 3; and Haack, *Deviant Logic, Fuzzy Logic*, pp. 229–58.

[29] See Haack, *Deviant Logic*, chapter 7.

[30] See Priest, *In Contradiction* (Boston: Martinus Nijhoff, 1987); Priest, *Introduction to Non-Classical Logic*, chapters 7–8; Priest, "The Logic of Paradox," *Journal of Philosophical Logic*, VIII (1979): 219–41; and Priest, "What Is So Bad about Contradictions?," this JOURNAL, XCV, 8 (August 1998): 410–426.

Perhaps the strongest argument in favor of the diehard classicist option is Quine's thesis that all linguistically competent human speakers find the logical constants, logical truths, and proof procedures of classical logic "obvious, actually or potentially" (*op. cit.*, p. 82) because they are intrinsic to the very practice of translation:

> Take the...case of trying to construe some unknown language on the strength of observable behavior. If a native is prepared to assent to some compound sentence but not to a constituent, this is a reason not to construe the construction as conjunction. If a native is prepared to assent to a constituent but not the compound, this is a reason not to construe the construction as alternation. We impute our orthodox logic to him, or impose it on him, by translating his language to suit. We build the logic into our manual of translation. Nor is there cause here for apology. We have to base translation on some kind of evidence, and what better (*op. cit.*, p. 82)?

It follows directly from this that if someone attempts to deny some fundamental law of elementary logic (say, the classical or strong law of noncontradiction[32]), then she cannot be intelligibly regarded as offering a genuine challenge to classical logic. If classical logic is built into our manual of translation, then the challenger must instead be regarded by us as meaning something different by the logical words she uses than we do by the very same words:

> What if someone were to reject the law of non-contradiction and so accept an occasional sentence and its negation both as true? An answer one hears is that this would vitiate all science. Any conjunction of the

---

[31] See, for example, Haack, *Deviant Logic*, p. 26.

[32] As I understand them, the strong laws of noncontradiction, bivalence, and excluded middle are as follows:

Strong Noncontradiction: For every sentence *S*, '*S* and not-*S*' is not true.
Strong Bivalence: For every sentence *S*, *S* is assigned one and no more than one of the two truth-values, true and false.

> form '$p \,.\, \sim p$' logically implies every sentence whatever; therefore acceptance of sentence and its negation as true would commit us to accepting every sentence as true, and thus forfeiting all distinction between true and false. In answer to this answer, one hears that such a full-width trivialization could perhaps be saved off by making compensatory adjustments to block this indiscriminate deducibility of all sentences from an inconsistency. Perhaps, it is suggested, we can so rig our new logic that it will isolate its contradictions and contain them. My view of this dialogue is that neither party knows what he is talking about. They think they are talking about negation, '~', 'not'; but surely the notation ceased to be recognizable as negation when they took to regarding some conjunctions of the form '$p \,.\, \sim p$' as true, and stopped regarding such sentences as implying all others. Here, evidently, is the deviant logician's predicament: when he tries to deny the doctrine he only changes the subject.[33]

Against Quine's diehard classicism, however, it seems to me that it cannot be seriously denied that both extended and deviant logic (even such highly deviant systems as dialetheic logic) are really and truly logic. This is because we must accept, I think, that any nonclassical logic (an NCL, for short) is really and truly logic just in case the following three conditions are satisfied:

(I) *The formal logic condition.* The NCL is a science of the necessary relation of consequence.

(II) *The representational adequacy condition.* The NCL's proposed extension of, or deviation from, classical logic is based on its being able accurately to represent in the format of symbolic logic some apparent linguistic facts that are not represented within classical logic: for example, strict implication or modality, constructibility of proofs, relevance, vagueness, future contingency, nonexistent objects, paradoxes, and so on.

(III) *The localization of application condition.* The NCL's scope of application is restricted to all and only those language domains containing the apparent nonclassical linguistic facts that it represents.

Strong Excluded Middle: For every sentence *S*, '*S* or not-*S*' is true.

It is possible to reject strong excluded middle without challenging strong bivalence (for example, intuitionism); to reject strong bivalence without challenging strong excluded middle (for example, dialetheism); to reject strong noncontradiction without rejecting strong excluded middle (for example, dialetheism again); and to reject strong bivalence without rejecting strong noncontradiction (for example, three-valued logic).

[33] Quine, p. 81.

Conditions (I) to (III) say, truistically, that nonclassical logics are logics that apply to the nonclassical linguistic domains they adequately represent. So nonclassical logic is really and truly logic, and diehard classicism is wrong. But diehard nonclassicism has the same problem in the reverse direction. It cannot be denied that classical or elementary logic is really and truly logic, since the three conditions of formal logic, representational adequacy, and localization of application are all trivially satisfied by elementary logic for apparent classical logical linguistic facts.

Are we then forced into the all-embracing arms of the unconstrained pluralist option? No. This is because it seems to me that Quine overstates his case in favor of diehard classicism, and that a slightly weaker version of his argument in fact strongly supports a distinct fourth option that is neither diehard classicism, nor diehard nonclassicism, nor unconstrained pluralism. This fourth option asserts that *neither* is it true that there is One True Logic, whether elementary logic or nonclassical logic, *nor* is it true that the several different logics are all equally acceptable.

Suppose then that we accept that *some one single logic* is "obvious, actually or potentially" because it is (among other things) built into the very practice of translation, but also that this logic is itself neither strictly speaking classical or elementary logic nor strictly speaking nonclassical logic (whether extended or deviant), because it is stucturally distinct from any classical or nonclassical logical system. Indeed, it is not a logical *system* as such, but instead a single set of *schematic logical structures*, in the form of a coherent repertoire of metalogical principles and logical concepts. Furthermore, it is presupposed by every logical system whatsoever. This is because it is a *protologic*, in the sense that it is used for the construction of every actual or possible logical system. Such a universally presupposed constructive logic would be *somewhat* like Quine's own "sheer logic," but with his diehard classicism subtracted out. More precisely, it would be *quite* like Shapiro's "super-logic," the logic that we use for the internal analysis and cross-logic evaluation of the plurality of logics, which is also unrevisable and a priori. The protologic, as I am conceiving it, is unrevisable and a priori precisely because its total set of schematic logical structures determines what will *count* as a possible logical system, and because some knowledge of this set of structures must also be consciously *available* to thinkers if they are to be able to justify assertions or claims made about any classical or nonclassical logic. So the protologic is both constructively and epistemically presupposed by every logical system.

This fundamental point requires more elaboration. To say that some single universal set of schematic structures is constructively presupposed by every member of a plurality of formal systems of some definite sort (say, mathematical systems or linguistic systems), does not in and of itself guarantee that this set of schematic structures is unrevisable and a priori. Take, for example, the crucially analogous case of natural language. The mere fact that, say, Chomsky's universal grammar is constructively presupposed by every natural language does not in and of itself guarantee that the universal grammar is consciously available to competent speakers of a natural language, whether in the form of noninferential beliefs (also called "intuitions" if they are either prima facie compelling or intrinsically compelling) or inferential beliefs. But the case of logic is crucially different. This is because competent thinkers must be able to (try to) *justify* the assertions they make, and logical principles and concepts are always more or less consciously explicit in the connections between reasons for believing and the beliefs based on those reasons. What I mean is that the "because" in "I believe that *Q*, because of the fact that *P* (or: because it seems to me very likely that *P*, and so on)" is necessarily a *logical* "because": it says that something (that is, that *Q*) *logically follows* from something else (that is, that *P*). So all justification is, to some extent, conscious logical justification. And this obviously also covers the specific case in which justification is concerned with conscious beliefs *about* logic. To justify any assertion is to invoke conscious logical beliefs; so to justify any assertion about logic is to invoke conscious logical beliefs about logic; but nothing will count as a logic unless it presupposes the protologic; therefore in order to justify any assertions about any logic, we must invoke conscious logical beliefs about the protologic. In this way the protologic is unrevisable and a priori because it is not only constructively but also epistemically presupposed by every logic.

If this proposal is correct, then the unconstrained pluralist option is also ruled out, because the single universally (constructively and epistemically) presupposed protologic is *not* on a par with any of the many logical systems. On the contrary, the single universally presupposed protologic is—to use Kantian language for a moment—*the condition of the possibility of there being informative natural language discourse (including of course translation) and reasoning (including of course theories) in the first place.*

Now of course the $64,000 question becomes: What is the precise structural description of the protologic? This, however, is a question that obviously cannot be answered within the scope of this essay. I will allow myself to admit here that it *seems* to me that the following four metalogical principles, together with the logical concepts implicit in them, are at least good *candidates* for belonging to the protologic:

(i) *The weak principle of validity*: an argument is valid if it is impossible for all of its premises to be true and its conclusion false.
(ii) *The weak principle of noncontradiction*: not every sentence is both true and false.[34]
(iii) *The weak principle of logical truth*: a sentence is logically true if it comes out true under every possible uniform reinterpretation of its nonlogical constants.
(iv) *The weak principle of the transfer from logical truth to valid proof*: a proof from a set of premises to a conclusion is valid if the corresponding classical conditional of its underlying argument is logically true.

Each of these principles is a weak or minimal version of a basic metalogical principle of classical or elementary logic. My rationale for tentatively proposing the inclusion of these weak or minimal classical metalogical principles in the protologic is the (to me) very plausible thought that while every extended or deviant variant on classical logic adds something to or subtracts something from classical logic, *no* logical system can reject absolutely *all* of classical logic and still *remain* a logic. This in turn would suggest that the protologic, among other things, captures a weak or minimal classical "core" that is preserved in every classical or nonclassical system. So the protologic would also to that extent capture the *core* of classical logic.[35] But that is as far as I am willing to go.

This is not an evasion; it is simply the scientific division of labor, and a little theoretical modesty too. My intention in this section so far has been to argue that a protologic must exist, and to spell out the basic features that the protologic will intrinsically possess: it is a single universal set of schematic logical structures, in the form of a repertoire of metalogical principles and logical concepts, that possesses an overall structure distinct from every classical or nonclassical logical system, and is used for the construction of every classical or nonclassical system; and it is unrevisable and a priori precisely because it is both constructively and epistemically presupposed by every logical system. But this ambitious claim having been (let us assume for a moment!) accepted, then there must then be a scientific division of labor. This is because the question of the precise structural description of the protologic is an intrinsically *formal-logical* question, not a *philosophical* question, and this essay—to the extent that it is about logic, as one of its two basic topics—deals only with philosophical logic and the philosophy of logic, and is *not* an essay in formal logical theory. Furthermore I am not a professional logician, so I am not competent to undertake this further investigation, even if I wanted to. It is also the case that if what I am going to argue in the last part of this section is sound, then the question of the precise structural description of the protologic is in part an irreducibly *empirical* question, in that it will depend on some factual results in the psycholinguistic part of cognitive psychology. These somewhat negative points should be regarded as carrying a definitely positive scientific upshot, however. For if I am correct, then the $64,000 question of the precise structural description of the protologic constitutes a new and important joint research program for logicians and cognitive psychologists.

---

[34] See also Hilary Putnam, "There Is at Least One A Priori Truth," in Putnam, *Realism and Reason: Philosophical Papers, Volume 3* (New York: Cambridge, 1983), pp. 98–114.

[35] See also Ken Warmbrōd, "Logical Constants," *Mind*, CVIII (1999): 503–38.

Let me suppose for the purposes of argument that the existence of the protologic has been been successfully established. Now I want to propose that the protologic is innately contained in a human cognitive faculty for logical representation, namely the logic faculty. And I want to argue for this claim empirically, by an extension of Chomsky's psycholinguistics.

According to Chomsky's psycholinguistics, whether in its early "transformational" version, or in its later and canonical "principles and parameters" version,[36] what is called a "phrase-structure" grammar (which is a grammar that decomposes all sentences into immediate constituents such as noun phrases, verb phrases, adjectival phrases, prepositional phrases, and so on, according to rules) applies to each natural language.[37] The phrase-structure grammar of each natural language differs significantly from every other. If, by analogy with classical or elementary logic, we think of the phrase-structure grammar of our own language as "classical" or "elementary," then necessarily every other natural-language grammar is either an "extension" of it (if it preserves the same set of phrase-structure rules as our language but adds more—as in idiolects and dialects) or else a "deviant" of it (if it lacks or modifies some of the phrase-structure rules of our language—as in foreign languages). Yet Chomsky thinks that all of these as it were classical or elementary, extended, and deviant phrase-structure grammars presuppose a single set of underlying principles and concepts. This set of principles and concepts constitutes a generative universal grammar (UG) that is realized in all particular natural language grammars, yet in different ways, by setting certain formal parameters differently for each natural language.

---

[36] See, for example, Chomsky, *Syntactic Structures* (The Hague: Mouton, 1957); Chomsky, "Review of B.F. Skinner's Verbal Behavior," *Language*, xxxv (1959), 26–58; Chomsky, *Aspects of the Theory of Syntax* (Cambridge: MIT, 1965); and Chomsky, *Knowledge of Language* (Westport, CT: Praeger, 1986).

[37] See, for example, Vivian Cook and Mark Newson, *Chomsky's Universal Grammar* (Cambridge: Blackwell, 1996, 2nd ed.); and John Lyons, *Introduction to Theoretical Linguistics* (New York: Cambridge, 1968).

Let me now apply this Chomskyan idea to the protologic. I have argued that the protologic consists in a single set of schematic logical structures, in the form of a coherent repertoire of metalogical principles and logical concepts, that is constructively presupposed by every logical system whatsoever, and is unrevisable and a priori (or epistemically presupposed). My proposal is that *the protologic stands to the many classical or elementary, extended, and deviant logics, precisely as Chomsky's UG stands to the many native, idiolectic or dialectic, and foreign natural languages.* It needs to be particularly re-emphasized that the protologic is *not* to be identified with classical or elementary logic, or indeed with any nonclassical logic, just as the UG is *not* to be identified with the grammar of one's native language, or indeed with the grammar of any other language. Therefore just as the UG is a super-grammar or sheer grammar that is nonempirically implicit in the construction, internal analysis, and cross-language comparison (which may be either consciously or nonconsciously available to speakers) of all natural languages, so too the protologic is a super-logic or sheer logic that is not merely nonempirically but also unrevisably implicit in the construction, internal analysis, and cross-logic comparison (which must be consciously as well as nonconsciously available to thinkers) of all classical or nonclassical logical systems.

According to Chomsky's view, cognitive mastery or knowledge of a natural language is a person's "linguistic competence," which expresses her possession of an innate, universally-shared "language faculty" within her total cognitive capacity:

> [O]ne of the faculties of the mind, common to the species, is a faculty of language that serves the two basic functions of rationalist theory: it provides a sensory system for the preliminary analysis of linguistic data, and a schematism that determines, quite narrowly, a certain class of grammars. Each grammar is a theory of a particular language, specifying formal and semantic properties of an infinite array of sentences. These sentences, each with its particular structure, constitute the language generated by the grammar. The languages so generated are those that can be "learned" in the normal way. The language faculty, given appro-

> priate stimulation, will construct a grammar; the person knows the language generated by the constructed grammar. This knowledge can then be used to understand what is heard and to produce discourse as an expression thought within the constraints of the internalized principles, in a manner appropriate to situations as these are conceived by other mental faculties, free of stimulus control.[38]

Furthermore, according to Chomsky, linguistic competence in this sense is to be sharply distinguished from "performance," or the actual use of language in concrete contexts. Performance is notoriously variable and partially determined by external, contingent factors. Competence, by contrast, consists in the construction and possession of a comprehensive mental representation of a natural language, which Chomsky also calls an "internalized language" or "I-language." The I-language is constructed by means of the innate "schematism" that is the generative UG:

> Let us define "universal grammar" (UG) as the system of principles, conditions, and rules that are elements or properties of all human languages not merely by accident but by necessity.... Thus UG can be taken as expressing "the essence of human language." UG will be invariant among humans. UG will specify what language learning must achieve, if it takes place successfully.... What is learned, the cognitive structure attained, must have the properties of UG, though it will have other properties as well, accidental properties. Each human language will conform to UG. If we were to construct a language violating UG, we would find that...it would not be learnable under normal conditions of access and exposure to data (*ibid.*, p. 29).

For the purposes of my argument in this section I am quite prepared to assume without further ado that Chomsky's psycholinguistics is on the whole well supported both philosophically and empirically. But what is *most* important for my account is the fact that Chomsky holds (in effect, if not precisely in name) that what I am calling the protologic is built right into the UG, hence built innately into our innate language faculty. As he puts it in the text quoted as one of the epigraphs for this essay:

> The logical notions are embedded in our deepest nature, in the very form of our language and thought, which is presumably why we can understand some kinds of logical systems quite readily, whereas others are inaccessible to us without considerable effort...if at all (*op. cit.*).

To avoid confusion, we must distinguish between (i) the most-deeply-cognitively-embedded logical notions being described by

---

[38] Chomsky, *Reflections on Language* (New York: Pantheon, 1975), pp. 12–13.

University Library

Chomsky here, which I am identifying with the innate protologic, and (ii) what Chomsky specifically calls "Logical Form" or LF[39] in the context of the current version of his psycholinguistic theory, "the minimalist program."[40] LF is a level of natural-language representation in Chomsky's current model of linguistic competence that (i) combines both the underlying logical form of the sentences of a given natural language and their underlying semantic interpretations, and (ii) is related to phonological structure by means of syntactic structure. Still, LF *presupposes* the most-deeply-cognitively-embedded logical notions, or what I am calling the innate protologic.

In any case, the Chomskyan idea of "logical notions [that] are embedded in our deepest nature, in the very form of our language and thought," or the idea (to use my twist on it) of the innate protologic, leads directly to the further idea of a "logical competence"[41] that is presupposed by linguistic competence. On this extended Chomskyan picture, every linguistically competent being constructs an internalized logic or "I-logic"[42] for the representation of the "natural logic"[43] of her own natural language, just insofar as she constructs an I-language for the representation of the grammar and semantics of her own natural language. In other words I am asserting that that every linguistically competent being constructs a *logic of thought* just insofar as she constructs a *language of thought.* But the crucial point for right now is that the innately-grounded logic of thought, as well as every other logical system whatsoever, both constructively and epistemically presupposes the protologic, just as any creature's innately-grounded I-language, as well as every other language whatsoever, constructively presupposes the UG.

This concludes my empirical argument for the logic faculty thesis, from Chomsky's psycholinguistics. Again, the nub of the argument is that on the assumption that Chomsky's psycholinguistics is factually true, then it entails the truth of the logic faculty thesis. This also means, however, that there is an irreducibly empirical component in the logic faculty thesis, and that the precise structural description of the protologic will depend in part on empirical results in cognitive psychology. So if I am correct, then there is a new and important joint research program for logicians and cognitive psychologists, built right on top of Chomskyan psycholinguistics, that consists in working out the precise structural description of the protologic.

---

[39] See Norbert Hornstein, *Logic as Grammar* (Cambridge: MIT, 1984); and Robert May, *Logical Form* (Cambridge: MIT, 1985).

[40] See, for example, Chomsky, *The Minimalist Programme* (Cambridge: MIT, 1995); and Chomsky, *New Horizons in the Study of Language and Mind* (New York: Cambridge, 2000).

[41] See John Macnamara, *A Border Dispute: The Place of Logic in Psychology* (Cambridge: MIT, 1986), chapter 2.

[42] Both the notion of an internalized logic and the term 'I-logic' are mine: but they obviously mimic Chomsky's "internalized language" or "I-language."

[43] A natural logic is the logic that expresses our actual reasoning procedures. See M. Braine, "On the Relation between the Natural Logic of Reasoning and Standard Logic," *Psychological Review,* LXXXV (1978): 1–21; and George Lakoff, "Linguistics and Natural Logic," in Gilbert Harman and Donald Davidson, eds., *Semantics of Natural Language,* (Boston: Reidel, 1972), pp. 545–665.

II. THE ETHICS OF LOGIC

By "the ethics of logic," I mean *the normativity of logic.* Normativity, as such, consists in the fact that there is a set of ideals, standards, guides, recommendations, commands, rules, principles, laws, and so on (hence "norms"), that govern human beliefs and intentional actions. Norms imply the existence of human values. Human values are, roughly speaking, whatever *matters* to human animals: whatever humans consciously care about, need, desire, prefer, choose, put their practical, affective, or cognitive faith in, and intentionally act for the sake of. As I will construe it, the normativity of something *X* is expressed by saying that there is something humans *ought to* (*or may*) *believe or do because of X.* In other words, the normativity of *X* is the role *X* plays in the giving of reasons for human belief or intentional action, that is, in the justification of human belief or intentional action. More precisely then, *X* is normative if and only if *X* can be directly cited as a reason for human belief or intentional action, or anyhow is intrinsic to some reason for human belief or intentional action.

In a logical context, normativity consists in a certain relation between logic and human reasoning. To say that logic is normative is to say that humans *ought* to reason soundly or validly (more generally, cogently). Otherwise put, the normativity of logic consists in the fact, if it is a fact, that the justification of human beliefs or intentional actions depends on our ability to reason cogently. *If it is a fact.* Closer inspection reveals that there are various philosophical options with respect to the normativity of logic.

The first and most basic set of options, a binary pair, concerns the fact (if it is a fact) of logic's normativity itself. The options here divide as to whether logic is either (1A) a *normative* (that is, prescriptive or evaluative) science, or else (1B) a *non-normative* (that is, descriptive or factual) science.

Assuming that logic is normative, there is a further binary pair of options that divide as to whether the normativity of logic is either (2A) an *intrinsic* (that is, a necessary, relational or nonrelational) feature of it, or else (2B) an *extrinsic* (that is, an accidental, relational or nonrelational) feature of it.[44]

A third binary pair of options also falls under the assumption that logic is normative, and divide as to whether logic has either (3A) *categorical* (that is, unconditional, noninstrumental) normativity, or else (3B) *hypothetical* (that is, conditional, instrumental) normativity. This distinction can be expressed by saying that something *X* is categorically normative if and only if humans ought to believe or do *Y* because of *X* under all sets of circumstances and primarily because of *X* alone; whereas something *X* is hypothetically normative if and only if humans ought to believe or do *Y* because of *X* only in certain circumstances and primarily because of something else *Z*.

The third pair combines with the second pair to provide a fourth and seemingly exhaustive set of options, all under the assumption that logic is normative. Thus logic is either (4A) *intrinsically categorically normative,* (4B) *extrinsically categorically normative,* (4C) *intrinsically hypothetically normative,* or (4D) *extrinsically hypothetically normative.* These options in turn map fairly smoothly onto various historically real philosophical doctrines about the nature of logic.

The general idea that logic and human reasoning or human thinking are normatively connected goes at least as far back as the Cartesian logicians of Port Royal in the seventeenth century. Arnaud and Nicole construed logic as *l'art de penser,* or the art of thinking.[45] According to this view, the nature of logic is that it tells us how we ought to reason or think if we want to be good metaphysicians, mathematicians, or natural scientists. Logic is strictly a systematic means to this end: in effect, a *technology* of thinking. So the Port Royalists held that logic is intrinsically hypothetically normative (and thus falls under (4C)). Essentially the same view was later defended in the mid-nineteenth century by Mill.[46] The logic-as-technology-of-thinking view, for reasons that I will mention shortly, went down in flames at the turn of the twentieth century along with logical psychologism, but has been resuscitated in a contemporary context by logical nonfactualists or expressivists, who hold that the nature of logic is prescriptive, instrumental, and social: the laws, truths, and proofs of logic are nothing but ways we ought to talk when logicizing, and we logicize in order to serve the ends of more basic human practices, such as belief formation.[47]

---

[44] For convenience, I focus on options that assume the normativity of logic. But strictly speaking, since something can be both intrinsically *non-F* and also extrinsically *F* (that is, in relation to something else), this opens up the possibility that logic is both intrinsically non-normative and also extrinsically normative. In fact, as we shall see, most recent and contemporary writers of introductory logic textbooks hold that logic is intrinsically non-normative and also extrinsically normative.

[45] See Arnauld and Nicole, *Logic or the Art of Thinking,* Jill Vance Buroker, trans. (New York: Cambridge, 1996); Stephen Gaukroger, *Cartesian Logic* (New York: Oxford, 1989); and William Kneale and Martha Kneale, *The Development of Logic* (New York: Oxford, 1984), pp. 315–20.

[46] See Mill, *System of Logic* (London: Longmans, Green, and Co., 1879); and Edmund Husserl, *Prolegomena to Pure Logic,* in *Logical Investigations,* Volumes 1–2 J.N. Findlay, trans. (London: Routledge and Kegan Paul, 1970), Volume 1, pp. 51–247, at p. 76.

The sharply different idea of logic as the science of "the laws of thought" goes back to Kant in the eighteenth century. It was later picked up by Boole in the mid-nineteenth, and again by Frege in the late nineteenth and early twentieth, for example:

> Like ethics, logic can also be called a normative science. How must I think in order to reach the goal, truth?... [T]he task we assign logic is only that of saying what holds with the utmost generality for all thinking, whatever its subject matter. We must assume that the rules for our thinking and for our holding something to be true are prescribed by the laws of truth.[48]

> [The laws of logic] have a special title to the name 'laws of thought' only if we mean to assert that they are the most general laws which prescribe the way in which one ought to think if one is to think at all.[49]

On the Kant-Boole-Frege view, logic is the universal, topic-neutral, a priori science of the necessary laws of truth, and also a pure normative science based directly on rationality itself. Logic tells us *how we ought to reason or think in every possible set of circumstances because this is required by the nature of rationality.* So logic is intrinsically categorically normative (and thus falls under (4A)): logic is a moral science. For convenience, I dub this the *moral science conception of logic.*

Ironically enough, despite Frege's explicit support for the moral science conception of logic, that conception did not survive the late nineteenth and early twentieth century critique of psychologism which Frege himself initiated. Not too surprisingly, this was mainly

---

[47] See, for example, Michael Resnik, "Logic: Normative or Descriptive? The Ethics of Belief or a Branch of Psychology?," *Philosophy of Science*, LII (1985): 221–38; and Resnik, "Ought There to Be But One Logic?," in B.J. Copeland, ed., *Logic and Reality: Essays on the Legacy of Arthur Prior* (New York: Oxford, 1996), pp. 489–517, at pp. 510–14.

[48] Frege, "Logic [1897]," in Frege, *Posthumous Writings*, P. Long et al., trans. (Chicago: University Press, 1979), pp. 127–51, at p. 128.

[49] Frege, *Basic Laws of Arithmetic* M. Furth trans. (Berkeley: California UP, 1964), p. 12.

Husserl's doing. In the *Prolegomena to Pure Logic,* Husserl vigorously rejected not only logical psychologism, but also the intrinsic normativity of logic. If sound, this rejection rules out the very ideas of both the intrinsic categorical normativity of logic (that is, (4A)) and also the intrinsic hypothetical normativity of logic (that is, (4C)), thus dispensing with Kant, Boole, Frege, Arnaud and Nicole, Mill, and contemporary logical nonfactualists in one fell swoop. More positively expressed, Husserl claims that logic is an intrinsically nonnormative, descriptive, or factual science (and thus falls under (1B)).

Following Husserl, most recent and contemporary writers of introductory logic textbooks start from the assumption that the intrinsic normativity of logic is as questionable as logical psychologism. So logic is an intrinsically non-normative, descriptive, or factual science. But at the same time they also hold that logic is *extrinsically* normative, in relation to everyday human reasoning.[50] According to these writers, we rational humans *ought* to follow logical rules, either because this is required by rationality itself (and rationality is extrinsic to logic) or because it promotes other important human interests. In other words, for these writers, despite the fact that logic is an intrinsically non-normative science, it remains either extrinsically categorically normative (and thus falls under (4B)) or extrinsically hypothetically normative (and thus falls under (4D)). In other words, they hold that despite the fact that logic is essentially the science of the necessary relation of consequence, or the science of proof, or the science of truth, or whatever, nevertheless it is a *very good thing indeed* to study logic.

This brings us to Otto Weininger. Weininger thinks that obeying the laws of logic is something we ought to do precisely because it belongs to the strict moral duties we have towards ourselves. He thereby holds that logic is intrinsically categorically normative (and thus falls under (4A)), but for reasons quite different from those found in the moral science conception of logic. According to the moral science conception, logic is intrinsically categorically normative because it is based on rationality itself (hence rationality is intrinsic to logic) and is also an *integral part of* human morality, namely the part that consists in justifying moral judgments and decisions, including direct moral arguments and reflective equilibrium. But Weininger takes the much stronger view that logic is *iden-*

---

[50] See, for example., Stephen Barker, *Elements of Logic* (New York: McGraw Hill, 1980, 3rd ed.), p. 4; Irving Copi, *Symbolic Logic* (New York: Macmillan, 1973, 4th ed.), p. 2; and Wesley Salmon, *Logic* (Englewood Cliffs, NJ: Prentice Hall, 1963), p. 8.

tical with human morality. Weininger's extreme view, in turn, seems to have been personally shared by the early Wittgenstein:

> [Wittgenstein] used to come to see me every evening at midnight, and pace up and down my room like a wild beast for three hours in agitated silence. Once I said to him: "Are you thinking about logic or about your sins?" "Both," he replied, and continued his pacing.[51]

In the rest of this section and the next, I sketch and defend a broadly Kantian and cognitivist version of the moral science conception of logic. As I do so however, it is crucial to remember that the moral science conception of logic and the highly moralistic Weininger-Wittgenstein conception of logic are distinct.

As I mentioned, Husserl's *Prolegomena* effectively prevented the transmission of Frege's moral science conception of logic into mainstream European and Anglo-American twentieth century philosophy, by attacking the thesis that logic is intrinsically normative. Husserl has three basic objections to the intrinsic normativity thesis.[52] First, logical laws, truths, and proofs are not framed by logicians in normative terms (for example, as imperatives), but rather purely in alethic modal terms (for example, as necessary laws, truths, or proofs). Second, every normative discipline presupposes a more basic theoretical discipline that establishes the existence and nature of the facts that are taken to have the relevant normative properties. So even if there are normative disciplines connected with logic, logic itself is needed as a distinct non-normative science in order to ground those normative disciplines. And third, the thesis that logic is normative entails logical psychologism, and logical psychologism is false: so it is also false that logic is intrinsically normative.

Husserl's first objection can be seen to fail as soon as we note that in natural deduction systems (for example, Gentzen's[53]), logical inference rules are explicitly expressed as Rylean "inference-tickets" or generalized permissions. And even in Frege's *Begriffsschrift*, which is an axiomatic system, inference rules are explicitly not accorded the status of truth bearers but instead are assigned a normative role.[54] So there is no inconsistency in ascribing alethic modal properties and normative properties to logical language within one and the same logical system. Indeed, since it is a general feature of logical consequence or entailment that every logical truth follows from any set of premises including the empty set of premises, then for every logical truth in every logical system there is a corresponding permission to infer that sentence from any set of premises. So for every logical truth in every logical system a corresponding normative sentence carrying the same logical force can be formulated.

---

[51] Russell, *Autobiography* (London: Unwin, 1975), p. 330.

[52] See Husserl, *Prolegomena to Pure Logic*, chapters 1–3.

[53] See Gerhard Gentzen, "Investigations into Logical Deduction," in *The Collected Papers of Gerhard Gentzen*, Mary Szabo, trans. (Amsterdam: North Holland, 1969), pp. 68–131.

[54] See Thomas Ricketts, "Frege, the *Tractatus*, and the Logocentric Predicament," *Philosophers' Annual*, VIII (1985): 247–59.

Husserl's second objection fails because it has a false assumption. He assumes that to claim that logic is intrinsically normative is automatically to make a reductive claim to the effect that logic is "nothing over and above" some set of normative facts. Assuming such a reduction, it would indeed follow that logic could be legitimately framed in alethic modal terms only if normative logical disciplines had a nonnormative logical grounding. Now it is true that some versions of the thesis that logic is intrinsically normative are indeed reductive. Most noncognitivist theories of logic, for example, can plausibly be read as reductive. But clearly it is also possible consistently to hold (1) that logic is a non-normative discipline in the sense that every logic explicitly contains within itself language that is intrinsically descriptive or factual, (2) that logic is *also* a normative discipline in the sense that every logic explicitly (as in natural deduction systems, and in some axiomatic systems) or implicitly (as in other axiomatic systems) contains within itself language that is intrinsically prescriptive or evaluative, and (3) that the nonnormative and normative parts of logic are complementary (so for every logical truth in every logical system, there is a corresponding normative sentence legitimating an inference to that sentence from any set of premises, and conversely) and mutually irreducible. On this nonreductive approach, logic is intrinsically descriptive *and* intrinsically prescriptive.

This leaves us with Husserl's third objection, to the effect that the intrinsic normativity of logic entails logical psychologism. And here is where the distinction between hypothetical and categorical normativity is crucially salient. Logical psychologism is a form of scientific naturalism that consists in the explanatory reduction of logic to empirical psychology.[55] Now Husserl makes the correct point that some forms of logical psychologism also assert the thesis that logic is intrinsically normative. Mill's theory of logic, for example, shows that it is possible to hold both that logic is intrinsically hypothetically normative and that logic is explanatorily reducible to empirical psychology. And it also seems to be true that if one wants to hold that logic is intrinsically normative *and also* that it is explanatorily reducible to empirical psychology, then one must also hold that logic is intrinsically *hypothetically* normative, because both require that logic is dependent on contingent facts of some sort, whether actual human interests or the natural facts. But the plain truth is that not every theory that takes logic to be intrinsically normative is psychologistic: at least some versions of the thesis that logic is intrinsically *categorically* normative are nonpsychologistic,[56] including Kant's and Frege's normative theories of logic. Take for example Kant's normative theory of logic. Kant holds that the logical "ought" has the same deontic force as the moral "ought":

---

[55] See Robert Hanna, *Rationality and Logic* (Cambridge: MIT, forthcoming), chapter 1.

> In logic...the question is not about *contingent* but about *necessary* rules; not how we think, but about how we ought to think.[57]
>
> What I call applied logic...is thus a representation of the understanding and the rules of its necessary use *in concreto*, namely under the contingent conditions of the subject.... General and pure logic is related to [applied logic] as pure morality, which contains the necessary moral laws of a free will in general, is related to the doctrine of virtue proper, which assesses these laws under the hindrances of the feelings, inclinations, and passions to which human beings are more or less subject.[58]

Kant's ethics, in turn, explicitly states that moral prescriptions *inherently fail* to be strictly determined by either actual human interests or the natural facts. It is built into the very idea of pure practical reason that the categorical imperative or moral law is universally binding on rational beings, including all rational humans; that the categorical imperative is underdetermined by all actual or possible human interests or natural facts; and that a rational agent is capable of acting not merely in accordance with but also *from* the categorical imperative and (if necessary) in opposition to all actual or possible human interests and all actual or possible natural facts.[59]

---

[56] It is also true that some versions of the thesis that logic is intrinsically hypothetically normative are non-psychologistic. See, for example, Resnik, "Logic: Normative or Descriptive? The Ethics of Belief or a Branch of Psychology?" Still, every psychologistic theory of logic holding that logic is intrinsically normative must also hold that logic is hypothetically normative.

[57] Kant, *The Jäsche Logic*, p. 529.

[58] Kant, *Critique of Pure Reason*, A54–55/B78–79.

[59] See Kant, *Groundwork of the Metaphysics of Morals*, in *Practical Philosophy*, Mary Gregor, trans. (New York: Cambridge, 1996), pp. 37–108.

So Kant's normative theory of logic entails that logical prescriptions are neither dependent on actual human interests nor logically strongly supervenient on the natural facts. That is, it entails that his logic is nonpsychologistic:

> Pure logic...has no empirical principles, thus it draws nothing from psychology.[60]

> Some logicians...do presuppose *psychological* principles in logic. But to bring such principles into logic is just as absurd as to derive morals from life.[61]

Kant's logic and ethics are of course controversial. The relevant point here, however, is not whether Kant's logic and ethics are *defensible,* but instead whether the thesis that logic is intrinsically categorically normative is *intelligible.* So Husserl's third objection does not apply to at least some views holding that logic is intrinsically categorically normative.

Rejecting Husserl's three objections to the intrinsic normativity of logic leaves open a window for Kantian constructivism in logical theory, in three ways. First, KCLT asserts that despite the fact that every classical or nonclassical logical system is a non-normative, factual, or descriptive science of the necessary relation of consequence, it *also* has a normative part corresponding directly to its non-normative part. This can be easily seen in natural deduction systems, which explicitly contain inference rules; and it is also implicit in the very idea of a logical truth, which corresponds to a permission to deduce it from any set of premises. But even more fundamentally this can be seen in the fact that the protologic, as a set of logical principles and concepts for constructing logical systems, is inherently normative precisely insofar as it is a set of schematic permissions to construct logical systems *in just these ways and not in other ways.* The protologic does not tell us how rational humans actually *do* construct logical systems under real-world conditions. Similarly, Chomsky's UG is a *prescriptive* and not a descriptive grammar: it tells us how humans are *permitted* to construct natural languages by virtue of their innate cognitive endowment for language, not how they actually *do* construct their natural languages under real-world conditions. Second, KCLT asserts that the non-normative and normative parts of logic are mutually complementary and mutually irreducible. Third, KCLT asserts that the protologic is intrinsically categorically normative for human reasoning: in other words, that the protologic supplies "basic a priori rules of rationality."[62] It is the third claim that is most important and most controversial. The task of the next section is to argue for it.

---

[60] Kant, *Critique of Pure Reason,* A54/B78.

[61] Kant, *The Jäsche Logic,* p. 529.

III. LOGIC AS A SYSTEM OF CATEGORICAL IMPERATIVES FOR REASONING

Before getting down to the argument itself however, I want to pursue a little further the parallel between KCLT on the one hand, and Kant's own logic and ethics on the other, as a way of elaborating the thesis that logic is intrinsically categorically normative. In her seminal essay, "Morality as a System of Hypothetical Imperatives," Philippa Foot directly and vigorously challenged all forms of Kantian ethics.[63] Reduced to its essentials, Foot argues that Kant's categorical imperative is an overly rigid, overly abstract, and ultimately empty bit of formalism: since it provides no motivation for following it, the categorical imperative fails to yield genuinely action-guiding principles. Universal prescriptivism of any sort fails. Only a system of hypothetical imperatives, that is, a system of practical commands or prescriptive rules based explicitly on natural or empirical facts about particular human needs and aims given under particular social-historical conditions, can genuinely guide action.

Morality, says Foot, is more akin to rules of etiquette than it is to a system of categorical prescriptions (*ibid.*, p. 164). Rules of etiquette are codes of conduct for highly localized and relatively unimportant human practices. The only difference between etiquette and morality is the extent to which members of the broader human community have achieved a certain wide solidarity or shared caring about the relevant practices. So morality is nothing but "optimally well-entrenched" or "globalized" etiquette, that is, etiquette that has established itself across all significant human practices, in the sense that all or most of the members of the human community have explicitly or tacitly adopted its otherwise merely local aims as "common aims." Given these common aims, and *only* given these common aims, a set of hypothetical, conditioned, or instrumental moral prescriptions will follow.

A precise analogue of Foot's complaint against Kantian ethics can be found in the conventionalist theory of logic. According to logical conventionalism, one can (whether individually or socially) create and adopt any logical system one likes by stipulating a certain set of logical axioms or postulates, and also the rules of formation, transformation, and interpretation. And it is a central feature of logical conventionalism that what motivates the formal system's creation and adoption is not itself cognitive or theoretical: indeed, it is essentially noncognitive or interest-driven, hence voluntaristic or at least pragmatic. This view is nicely encapsulated in Rudolf Carnap's famous declaration in section 17 of *The Logical Syntax of Language*:

---

[62] See Ralph Wedgwood, "The A Priori Rules of Rationality," *Philosophy and Phenomenological Research*, LIX (1999): 113–31, especially at p. 128.

[63] Foot, "Morality as a System of Hypothetical Imperatives," in *Virtues and Vices* (Berkeley: California UP, 1978), pp. 157–73.

> In the foregoing we have discussed several examples of negative requirements...by which certain common forms of language—methods of expression and of inference—would be excluded. Our attitude to requirements of this kind is given a general formulation in the *Principle of Tolerance:It is not our business to set up prohibitions, but to arrive at conventions.... In logic there are no morals.* Everyone is at liberty to build up his own logic, i.e., his own form of language, as he wishes (*op.cit.*, pp. 51–52).

The very idea of an intrinsically categorically normative logic is undermined, Carnap thinks, by the existence of nonclassical logics. We must be radically open to the unlimited possibilities for creating alternative logics and to the diverse human motivations for modifying and challenging classical logic. In other words for Carnap, at least where logic is concerned, "anything goes."

But there is an obvious way in which Carnap's conventionalism fails. As Quine famously argued, in order to give a conventionalist definition of logical truth, conventionalism must presuppose and use preconventionalized logic, and thereby logically stultifies itself.[64] Foot's conventionalist conception of morality suffers from an analogous problem. If morality is to be interestingly different from mere etiquette, that is, if one is to crank up a local system of hypothetical imperatives into a truly global system that applies even to *prospective* and *possible* members of our community, then one must presuppose and use the very idea of a categorical imperative. For in terms of its syntactic or semantic structure a "globalized" hypothetical imperative is nothing but a *relativized* categorical imperative. That is, a globalized hypothetical imperative to do A is nothing but a categorical imperative ("Everyone ought to do A") containing within its scope a material antecedent condition which is an arbitrarily-chosen set of *human interest postulates*, call that set "Γ":

Everyone ought (If Γ, then to do A).

---

[64] See Quine, "Truth by Convention," in *The Ways of Paradox* (Cambridge: Harvard, 1976, 2nd ed.), pp. 77–106.

Similarly, a hypothetically or relatively logically necessary sentence *H* is nothing but a relativized logical necessity. That is, *H* is nothing but the consequent of a logically necessary material conditional containing within its scope an antecedent material condition which is an arbitrarily-chosen set of theoretical postulates (for example, meaning-postulates, scientific laws, and so on), call it "Δ":

Logically necessarily $(\Delta \rightarrow H)$.

Here is the moral of the story. Just as you cannot rationally escape from preconventionalized logic by making logic conventional at the level of your object-language (since preconventionalized logic inevitably returns to haunt you in your metalanguage), so too you cannot rationally escape from categorical imperatives by making morality hypothetical or conventional at the level of first-order ethics (since categorical, preconventionalized morality inevitably returns to haunt you in your metaethics). And in a long footnote added to "Morality as System of Hypothetical Imperatives" in 1977, Foot as much as admits this:

> Kant's thought seems to be that universal rules are universally valid in that they are inescapable, that no one can contract out of morality, and above all that no one can say that as he does not happen to care about the ends of morality, morality does not apply to him. This thought about inescapability is very important, and we should pause to consider it. It is perhaps Kant's most compelling argument against the hypothetical imperative, and the one that may make Kantians of us all (*op. cit.*, p. 171).

And the same point goes, I hasten to add, for logic. To say that logic is intrinsically categorically normative is simply to say that logic is rationally humanly inescapable, or at least to say that *the protologic* is rationally humanly inescapable (as justification for this claim, see the argument for the intrinsic categorical normativity of the protologic to follow shortly).

In any case Foot seriously but instructively misinterprets Kantian ethics, as Onora O'Neill has shown.[65] It is a mistake to think of Kant's categorical imperative as a superstrong first-order principle for action (or what Kant calls a "maxim"), that is, as an all-purpose practical decision procedure or algorithm. On the contrary, the categorical imperative is a second-order procedural principle applying universally to maxims. Negatively described, the categorical imperative is a filter for screening out bad maxims; positively described, it is a constructive protocol for correctly generating maxims, given the multifarious array of concrete input materials to practical reasoning, including beliefs, desires, habits, personal situation, social-historical context, and so on. Thus the categorical imperative says, roughly:

---

[65] See Onora (Nell) O'Neill, *Acting on Principle* (New York: Columbia, 1975); and O'Neill, *Constructions of Reason* (New York: Cambridge, 1989).

> Act *only* according to those maxims that every rational human being could adopt, and that remain consistent with our innate rational capacity for constructing and acting upon maxims.

This version of Kantian ethics is what O'Neill, extending and modifying the work of John Rawls, aptly calls "Kantian constructivism."[66] The crucial point is that we cannot say in advance of actual practical reasoning processes just *which* maxims will turn out to be permissible or obligatory, but we can know a priori that any maxim that will count as action-guiding *must* have a format or structure that is determined by the categorical imperative.

Where this line of argument is heading should be fairly obvious. Just as a constructivist Kantian ethics can get around both the objections and the internal problems of moral conventionalism, so too KCLT can get around both the objections and the internal problems of logical conventionalism. And O'Neill even supplies us with a good map of the conceptual terrain I want to occupy:

> The Categorical Imperative is the supreme principle of reasoning not because it is an algorithm either for thought or for action, but because it is an indispensable strategy for disciplining thinking or action in ways that are not contingent on specific and variable circumstances. The Categorical Imperative is a fundamental strategy, not an algorithm; it is the fundamental strategy not just of morality but *of all activity that counts as reasoned.* The supreme principle of reason is merely the principle of thinking and acting on principles that can (not "do") hold for all.[67]

O'Neill is saying that the basis for the construction of any rational scheme of principles, whether that scheme is to be thought-guiding (logic) or intentional action-guiding (morality), is the categorical imperative. As applied to intentional action, the categorical imperative says that any first-order action-guiding principle must be universalizable, nonexploitative, and so on. But as a KCLT theorist would put it, as applied to thought, the categorical imperative says that every reasoning process must conform to the protologic:

---

[66] See O'Neill, *Constructions of Reason*, chapter 11. See also Thomas Hill, "Kantian Constructivism in Ethics," *Ethics*, XCIX (1989): 752–70; Rawls, "Kantian Constructivism in Moral Theory"; and Rawls, "Themes in Kant's Moral Philosophy," in Eckhart Förster, ed., *Kant's Transcendental Deductions* (Stanford: University Press, 1989), pp. 81–113.

[67] O'Neill, *Constructions of Reason*, pp. 58–59, emphasis added.

> Think *only* according to those processes of reasoning that satisfy the protologic.

Thus according to KCLT, following the principles and concepts of the protologic is not only required by human reasoning, but is also a *strict duty* for all human reasoning. And even though any such process of reasoning constructed in our logic of thought and our language of thought by means of the protologic will always have a more or less limited application that is determined by the inescapably contingent creature-based and world-based constraints under which that reasoning process occurs, *nevertheless* reasoning according to the principles and concepts of the protologic remains an obligation for every rational human animal in every cognitive context whatsoever.

In this way, to say that "logic is a system of categorical imperatives for reasoning" is just to say that *the protologic is intrinsically categorically normative for human reasoning.* Here now is an explicit argument for this thesis.

(1) Something is normative if and only if it can be cited as a reason for human belief or intentional action, or is intrinsic to some reason for human belief or intentional action. (From the definition of the concept of "normativity" in section I.)
(2) Every inferential justification of human belief or intentional action involves the logical entailment of some sentence describing a belief (call it a "belief report") or an intentional action (call it an "action-report") by premises that describe reasons for that belief or action (call them "belief premises" or "action premises" respectively). (Premise.)
(3) The logical entailment of a belief report or action report by belief premises or action premises, as understood by a rational human agent, involves some or another concept of logical consequence in the agent's logic of thought. (From (2), and the logic faculty thesis.)
(4) The protologic enters intrinsically into every logical system insofar as it is both constructively and epistemically presupposed by every logical system. (From the logical faculty thesis.)
(5) So the protologic enters intrinsically into every inferential justification of human belief or intentional action. (From (2) – (4).)
(6) So the protologic is intrinsically normative. (From (1) and (5).)
(7) The intrinsic normativity of the protologic is not instrumental, or based on actual human interests, because the protologic is not dependent on contingent facts, including of course actual human interests. (From the logic faculty thesis.)

(8) Therefore the protologic is intrinsically categorically normative. (From (6) and (7).)

(9) The protologic enters intrinsically into all human reasoning, because (i) the protologic is used and presupposed in the construction, self-analysis, and self-evaluation of each rational human cognizer's own logic of thought, and (ii) each rational human cognizer's logic of thought constitutes his processes of reasoning. (From the logic faculty thesis.)

(10) Therefore the protologic is intrinsically categorically normative for human reasoning. (From (8) and (9).)

As I indicated in steps (1), (3), (4), (7), and (9), this argument relies heavily on the logic faculty thesis, which I have argued for in section I and need not defend again here. The only step that needs independent support is step (2). Here the simple rationale is that if I am inferentially to justify a belief of mine or an intentional action of mine then obviously I must offer reasons for that belief or intentional action, and just as obviously the premises describing those reasons must *logically entail*, under some or another version of logical consequence or entailment, a conclusion describing my belief or my intentional action. How *else* could I inferentially justify a belief or intentional action of mine? Inferential justification, or the inferential giving of sufficient reasons for some human belief or intentional action (and this captures the very pith and marrow of a rational "because") is obviously inherently *logical*, precisely because it is *inferential*.

It should be noted that the version of logical consequence or entailment by virtue of which my belief premises or action premises entail my belief report or action report, need not always be *classical* consequence or entailment (that is, the sort of consequence or entailment we find in elementary logic), but can in principle be any sort of classical or nonclassical consequence or entailment. KCLT does not say that any *particular* classical or nonclassical logic supplies categorical imperatives for all human reasoning. Instead KCLT says that any logical system we use in any sort of reasoning more or less implicitly, but always intrinsically, invokes the principles and concepts of the protologic as categorical imperatives for that reasoning.

IV. CONCLUSION

You have probably already thought of an obvious objection to my claim, which follows directly from my Kantian constructivism in logical theory, that rational human animals are inherently logical animals in the sense that they are logical moralists. The objection is that I have falsely substituted an ideally rational, perfect, unhuman, Mr Spock-like (or Data-like, for younger Trekkies) reasoner for real, imperfect, rational *human animals*. But we all know from first-hand experience that we constantly reason inconsistently or invalidly, shift to mere rhetoric and sophistry when moved by our desires or emotions, get confused, and so on. And the well-known empirical research in cognitive psychology on "the reasoning tests," and in particular on "Wason's selection task," shows clearly that humans of ordinary or higher intellectual ability are generally very bad at logical deduction tasks.[68] So you will say that we cannot possibly be inherently logical animals in the sense that we are logical moralists: How then could we ever make a logical mistake?

My reply is that the obvious objection is based on an error about how the word 'logical' functions in "logical animals." 'Logical' as it used in this phrase is not primarily a descriptive term referring to any of the sciences of the necessary relation of consequence, that is, to any particular classical or nonclassical logical system, but instead is primarily a *prescriptive* term implicitly referring to the protologic. The protologic is intrinsically categorically normative for us, and there are good reasons to hold that we do indeed perfectly obey the protologic whenever we reason:[69] but it certainly does *not* follow that we always or even usually reason cogently under real-world conditions in relation to any particular classical or nonclassical logical system. As Boole very aptly remarks:

> The...laws of reasoning are, properly speaking, the laws of right reasoning only, and their actual transgression is a perpetually recurring phenomenon. Error, which has no place in the material system [of physical nature], occupies a large one one here. We must accept this as one of those ultimate facts.[70]
>
> Objection that we do not actually reason thus. Reply: It is a mistake to suppose that the actual performances of our nature in any case fully

---

[68] See Peter Wason, "Reasoning," in Brian Foss, ed., *New Horizons in Psychology* (Baltimore: Penguin, 1966), pp. 135–51; Wason, "Reasoning about a Rule," *Quarterly Journal of Experimental Psychology*, xx (1968): 273–81; and Wason and Philip Johnson-Laird, *The Psychology of Reasoning* (Cambridge: Harvard, 1972). Similar results have been discovered in connection with probabilistic reasoning tasks. See Daniel Kahneman, Paul Slovic, and Amis Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (New York: Cambridge, 1982); and Tversky and Kahneman, "Extensional vs. Intuitive Reasoning: The Conjunctive Fallacy in Probability Judgment," *Psychology Review*, xc (1983): 293–315.

[69] See Hanna, chapter 5.

[70] Boole, *An Investigation of the Laws of Thought* (Cambridge: Macmillan, 1854), p. 408.

> answer to its faculties and capacities. We are in all things constituted with reference to an ideal standard.[71]

In other words, rational human animals are creatures with very high logical standards in one sense (as regards our conception of an ideal logical reasoner); and with very high logical success rates in another sense (as regards the protologic); but, sadly, also with very low logical success rates in still another sense (as regards any particular constructed logical system, classical or nonclassical, given the multitude of nonlogical factors affecting logical performance in the real world[72]). The categorically normative "ought" governing logical reasoning performance certainly does not imply a factual "is." The crucial general fact about our rationality in respect of logic is not our logical performance, but rather that reasoning cogently inevitably *matters* to us. Just as only an essentially moral animal would ever care about committing sins, so too only an essentially logical animal would ever care about committing fallacies. But just because by virtue of our being inherently moral and logical animals, it follows that we rational humans thereby care both constitutively and profoundly about morality and logic, that does not stop us from committing *lots and lots* of sins and fallacies. Alas.

ROBERT HANNA

University of Colorado, Boulder

---

[71] Boole, "Extracts from a Paper Entitled 'On the Mathematical Theory of Logic and on the Philosophical Interpretations of Its Methods and Processes' (1855 or 1856)," in Boole, *Studies in Logic and Probability*, pp. 230–46, at p. 246.

[72] See, for example, Richard Griggs and James Cox, "The Elusive Thematic Materials Effect in Wason's Selection Task," *British Journal of Psychology*, LXXIII (1982): 407–20.

COMMENTS AND CRITICISM

BAYS, STEINER, AND WITTGENSTEIN'S "NOTORIOUS" PARAGRAPH ABOUT THE GÖDEL THEOREM

How would it be if *Principia Mathematica* (hereafter PM) turned out to be ω-inconsistent?[1] In our "A Note on Wittgenstein's 'Notorious Paragraph' about the Gödel Theorem,"[2] we showed that in 1937, when he wrote down the most "notorious" of his remarks about Gödel, Wittgenstein may well have been pondering this question.[3] Viewed in this light, we argued, his remarks may be seen to contain the germ of a significant philosophical insight about Gödel's theorem, rather than a hopeless effort to refute it. The purpose of our "Note" was to characterize this insight, detaching it from the disputed question whether Wittgenstein fully understood the Gödel theorem.

In a reply, Timothy Bays alleges that we have given a philosophical answer to the question "How would it be if PM turned out to be ω-inconsistent?" that is "almost certainly false" based upon an argument that is "inadequate."[4] According to Bays our aim was, first, to urge upon our readers the abandonment of ℕ, the standard model of arithmetic (204 n.19) and, second, to deny what Bays takes to be clearly true, namely, that there are true but unprovable propositions of arithmetic (201ff.).

As we shall explain, Bays misconstrues our claims. Moreover, he forwards a view of the notion of truth that we find problematic.

What did we actually claim? Wittgenstein's insight, as we construed it, asks us to appreciate the philosophical naiveté and/or metaphys-

---

[1] A formal system L is *ω-inconsistent* if there exists some well-formed formula $P(v)$, expressing a predicate of natural numbers and with no free variable other than v, such that $(\exists v)Pv$ is provable in L and yet so are all the formulas $\neg P(\underline{0})$, $\neg P(\underline{1})$, $\neg P(\underline{2})$, ... (where $\underline{0}$, $\underline{1}$, $\underline{2}$, ...are the formal expressions for the natural numbers in L). L is *ω-consistent* if there is no such well-formed formula.

[2] Juliet Floyd and Hilary Putnam, "A Note on Wittgenstein's 'Notorious Paragraph' about the Gödel Theorem," in this JOURNAL, XCVII, 11 (November 2000): 624–32; hereafter FP.

[3] The "notorious" remarks, written in the fall of 1937, are published in Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics* (Cambridge: MIT, 1978, revised edition), Part I Appendix III §8.

[4] Timothy Bays, "On Floyd and Putnam on Wittgenstein on Gödel," this JOURNAL, CI, 4 (April 2004): 197–210.

0022-362X/06/0302/101–10

ical unclarity involved in taking the following claim to be a straightforward truth:

> *Claim:*
>
> Gödel's theorem proves that
>
> (1) there is a well-defined notion of "mathematical truth" applicable to every formula of *Principia Mathematica*
>
> and
>
> (2) if *Principia Mathematica* is consistent, then some "mathematical truths" in the sense of (1) above are undecidable in PM.

Inspired by Wittgenstein's notorious remarks, we argued as follows. Suppose that to our surprise we have discovered a proof of the negation of a Gödel sentence, "$\neg P$", in *Principia Mathematica.*[5] Suppose too that PM is consistent. Under these suppositions it is a well-known, uncontroversial consequence of Gödel's theorem that PM will be ω-inconsistent. As may be seen from inspection of the definition of "ω-inconsistent" (see footnote 1), this means that every model of PM must contain entities that are not natural numbers. In fact, in any such model *every* numerical predicate with an infinite extension will "overspill," that is, contain some elements that are not natural numbers. But then, because our original rigorization of the syntactic notions applied in the English sentence came through Gödel numbering with the natural numbers alone (as in Gödel's original (1931) paper[6]), our initial way of translating into English or German P (the formula of PM whose proposed translation was "P is not provable in PM") would, in this context, become more nuanced and complicated, would have to be "given up" in its original form, just as Wittgenstein observed.

Despite what Bays says (198, 201, 202–03, 208 n.27), neither we nor Wittgenstein as we understand him were trying to argue that there are *no* circumstances under which we may legitimately translate P into the English sentence "P is not provable" (or say that P itself is "true but unprovable"). Indeed, our interpretation (unlike Bays's own, which we shall discuss below) prescinds altogether from talk about all proper or possible translations or interpretations of *P*. Instead, we take Wittgenstein to have been exploring the context-dependent *sense* of the claim that "there are true but unprovable sentences of arithmetic" as it applies both to PM and to our natural language, and doing so within a particular philosophical and historical context. We did connect Wittgenstein's remark about "giving up" an interpretation of *P* with his (commendable) rejection of a notion of *interpretation* which in principle wholly depends upon formalized mathematical language: as we wrote, Wittgenstein was "denying that a formal system *could* provide us with a standard of truth or clarity that is, in principle, inaccessible to a natural language" (FP 632). Of course, the argument in our "Note" itself then turns on a partly informal understanding of the notion a sentence's being "true in an interpretation" (or "satisfied in a model"). As we all agree, in 1937 model theory was not yet an established branch of mathematics, and Wittgenstein may be forgiven for leaving it out of his purview. And we believe that the later development of formal semantics and model theory fail to impugn the importance of Wittgenstein's philosophical insight about the Gödel theorem.

---

[5] P has the form: $\neg(\exists x)(\text{NaturalNo.}(x).\text{Proof}(x,t))$, where "$t$" abbreviates a numerical expression whose value calculates out to be the Gödel number of P itself, "Proof" abbreviates a predicate which is supposed to define an effectively calculable relation which holds between two natural numbers n,m just in case n is the Gödel number of a proof whose last line is the formula with Gödel number m, and "NaturalNo.(x)" is the predicate of PM we interpret as "x is a natural number."

[6] *Kurt Gödel: Collected Works, Volume I*, Solomon Feferman et al., eds. (New York: Oxford, 1986), pp. 145–95.

Bays thinks that our interpretation works against conceiving the standard model of arithmetic, ℕ, Th(ℕ)—and even model theory generally—as viable objects of study for mathematicians. He alleges that our "Note" leads ineluctably to "the abandonment of ℕ that Floyd and Putnam urge upon us" (204 n.19), to a principled rejection of "the semantical analysis of formal systems," to a "crazy" position which is offering "principled reasons for permitting us to step back to make syntactic generalizations, while forbidding us to step back to make semantic generalizations" (210, cf. n.31). But the reader will see no trace of these ideas, either in our "Note" or in the argument sketched above.

So what has happened?

Although Bays says he will make no interpretative claims about Wittgenstein (197), he relies on several assumptions about the notorious remarks and our interpretation of them that are, unfortunately, mistaken. The first is his idea that, inspired by Wittgenstein, we wrote about Gödel in the grip of anti-realism, verificationism, and/or formalism, and that we identified a priori the notions of "mathematically true" and "proved" (or perhaps, less stringently, "provable" or "provable in some formal system or other"). It does follow immediately from this identification that there can be no true but unprovable sentences of arithmetic. But contrary to what Bays says (202 n.14), the argument presented in our "Note" neither depends upon nor was used to advocate such a philosophical point of view. Nor does it depend upon the assumption, which Bays also attributes to us, that Wittgenstein is "opposed in principle to the derivation of 'philosophical' claims from 'mathematical' arguments" (198). We were arguing, instead, that neither formalism nor verificationism about mathematical truth in general are at issue in Wittgenstein's 1937 remarks on Gödel.[7]

Moreover, Bays has taken for granted a central strand of Mark Steiner's recent very interesting interpretation and criticism of Wittgenstein's notorious remarks.[8] Now Steiner himself rejects, as we do, the idea that Wittgenstein was a verificationist.[9] But he does present a reading that is starkly at odds with ours.[10]

According to the Steiner-Bays interpretation of the notorious remarks, Wittgenstein aspired to show that the claim that "there are true but unprovable sentences of arithmetic" is false, nothing but a misguided philosophical interpretation of the theorem. Bays argues (following Steiner) that this argument must be wrong. For, as he explains, as a result of Tarski's analysis of the notion of a sentence being *true in a formalized language* (or *satisfied in a structure*) we can give a proof in formal semantics of the Gödel theorem that demonstrates the existence of "true but unprovable" sentences of arithmetic. As Steiner has written, "it is a *mathematical theorem* (of set theory) that the Gödel sentence P is true in Tarski's sense if and only if it is not provable in Peano Arithmetic."[11] This implies that if the Gödel sentence P were false, it would be provable in Peano Arithmetic

---

[7] Several reasons for doubting that the notorious remarks express verificationism about mathematical truth as a whole are canvassed in Floyd, "On Saying What You Really Want to Say: Wittgenstein, Gödel and the Trisection of the Angle," in Jaakko Hintikka, ed., *From Dedekind to Gödel: The Foundations of Mathematics in the Early Twentieth Century* (Boston: Kluwer, 1995), pp. 373–426. For the larger issue of verificationist tendencies in Wittgenstein's other later remarks, see Putnam, "Was Wittgenstein *Really* an Antirealist about Mathematics?" in Timothy G. McCarthy and Sean Stidd, eds., *Wittgenstein in America* (New York: Oxford, 2001), pp. 140–94.

[8] Mark Steiner, "Wittgenstein as His Own Worst Enemy: The Case of Gödel's Theorem," *Philosophia Mathematica*, IX (2001): 257–79.

[9] Personal correspondence; compare Steiner's "Wittgenstein: Mathematics, Regularities and Rules," in Adam Morton and Stephen Stich, eds., *Benacerraf and His Critics* (Cambridge: Blackwell, 1996), pp. 190–212.

[10] A survey of the issues raised by Steiner's own rather sophisticated understanding of Wittgenstein on Gödel lies outside the scope of this reply, but some of them have been discussed in Floyd, "Prose versus Proof: Wittgenstein on Gödel, Tarski and Truth," *Philosophia Mathematica*, III, 9 (2001): 901–28, and in Putnam, "A Note on Steiner on Wittgenstein, Gödel, and Tarski" (*Iyyun*, forthcoming).

[11] Steiner, "Wittgenstein as His Own Worst Enemy," p. 267.

(hereafter PA), and we would then have a false theorem of PA. But, Bays argues (again following Steiner), this cannot be, because we have another mathematical theorem that all theorems of PA are true (in Tarski's sense). In other words, we may safely assume that PA is sound, and this implies that there are true but unprovable sentences of arithmetic (200).

This interpretation of Wittgenstein on Gödel avoids engaging with the issue of ω-inconsistency altogether, something neither Gödel's original paper nor Wittgenstein's remarks on that paper (as we read them) do. For it relies on the standard contemporary exposition of the Gödel theorem using an improved version of Gödel's proof due to J.B. Rosser.[12] Gödel showed that if PM is consistent, then P is not provable, but he did not see how to show that if PM is consistent, then $\neg P$ is not provable. Gödel realized that the latter argument requires the stronger assumption that PM is ω-consistent—so his proof was not "symmetric" about P and $\neg P$. In 1936 (in a paper that so far as we know Wittgenstein never saw) Rosser showed that this asymmetry, and Gödel's stronger assumption of ω-consistency, are not necessary for a proof of the incompleteness of PM.[13] For we can construct a sentence of PM, call it "R", such that a proof of R in PM can be converted algorithmically into a proof of $\neg R$, *and vice versa*, assuming only the simple consistency of PM. Freely translated, R expresses "for all n, if n is the Gödel number of a proof of R, then there is a proof of $\neg R$ with a Gödel number smaller than n," or, even more freely, "If R is provable, then $\neg R$ is provable even more quickly."

Observe, however, that R was not the sentence P which either Wittgenstein or our "Note" were discussing! Bays's shift to a new context—one that dispenses altogether with the sentence of PM Wittgenstein and Gödel were contemplating—leaves behind the whole issue of a sentence that *purports* to say of itself that it is true but unprovable, *full stop.* Since that was the sentence and the issue in which Gödel and Wittgenstein were, as we see it, interested, Bays's idealization fails to engage with the philosophical claim we take Wittgenstein to have been investigating in his remark.

Bays's presentation of the Gödel theorem also (like Steiner's) substitutes PA for PM in considering the application of Gödel's theorem to the case in which we imagine that $\neg P$ is demonstrated—

---

[12] So did Steiner, "Wittgenstein as His Own Worst Enemy," pp. 259, 262.

[13] Rosser, "Extensions of Some Theorems of Gödel and Church," *The Journal of Symbolic Logic*, I (1936): 87–91.

as Bays explicitly acknowledges. But this substitution also dodges the issue we take Wittgenstein to have been raising. Bays's assumption is that the soundness of PA can be proved in whatever system we employ to formalize Tarski's theory of truth. But that is not the same as assuming that that system can prove the soundness of Russell's *Principia Mathematica*! If we are right in our understanding of what Wittgenstein has in mind in writing "suppose the opposite is proved in Russell's system," then his whole point was to ask whether we would hold on to our English interpretation of P as "P is not provable" if $\neg P$ were proved *and we therefore realized that PM was not sound.* There is no reason to think that Wittgenstein would have regarded PA in the same light as PM—and no reason to think that we, looking back all these decades later, should do so either (as Bays himself says (207 n.22)).

This is not merely an historical aside, it is a logical point. Bays's criticism of the argument he attributes to Wittgenstein (like Steiner's) ultimately turns on invoking a semantical proof of the Gödel theorem that is a nonconstructive proof in set theory. Tarski himself did not actually give such a proof, but his definition of *satisfaction-in-a-model* does yield the means to express in set theory, not only the notion of a sentence being *true in arithmetic,* but also the idea of a set being first-order *definable* in a formalized language of arithmetic.[14] Putting this together with Gödel's manner of defining the notion of *provable in arithmetic* (*via* Gödel numbering), we can then express the following argument in set theory:

> *Truth in arithmetic* is not definable in a formalized theory of arithmetic, but, as Gödel showed, *provability in arithmetic* is. Hence the two notions cannot be coextensive.

Gödel claimed to have already grasped this (nonconstructive) argument in 1930, but he did not publish it, partly because of the finitistic context of the Hilbert Program, and partly out of worry that philosophers hostile to the notion of *truth* might object to his result.[15] What Gödel's 1931 paper (the one Wittgenstein saw!) gives, quite self-consciously, is a different, finitistically acceptable argument for the incompleteness of arithmetic *via* an explicit construction of an undecidable sentence. As Gödel wrote to Menger after reading Wittgenstein's notorious remarks about his theorem, his incompleteness result "is a mathematical theorem within an absolutely uncontroversial part of mathematics (finitary number theory or combinatorics)."[16]

---

[14] For a first-order language L, a structure $\mathbb{M}$ interpreting L, a formula ψ of L with k free variables and $a_1,\ldots,a_k$ elements of the domain M of $\mathbb{M}$, "$\mathbb{M} \models \psi\ [a_1,\ldots,a_k]$" is taken to assert that the formula ψ is satisfied in $\mathbb{M}$ when the k free variables are assigned sequentially to $a_1,\ldots,a_k$. Tarski's definition of satisfaction in a structure (or model), his analysis of "truth in an interpretation," amounts to a precise, set-theoretic definition of this notion. Applied to the particular structure of arithmetic that model theorists call "$\mathbb{N}$" (to be characterized in the final section of our essay, below) it provides an analysis of the notion of a sentence being *true in arithmetic.* Now with L and $\mathbb{M}$ as above, a set $X \subseteq M$ is *definable in L* if there is a formula ψ in $n+1$ free variables and "parameters" $b_1,\ldots,b_n$ such that $X = \{a \in M \mid \mathbb{M} \models \psi\ [a, b_1,\ldots,b_n]\}$.

If we assume that arithmetic is bivalent[17] and look at the Gödel theorem not through the eyes of Rosser's improvement, but with Wittgenstein's point in mind, then we can say that what the Gödel theorem shows is *either* that (1) there exists a true but unprovable sentence of number theory in PM *or* that (2) number theory is inexpressible in PM. For if PM is consistent, and yet we had a proof that $\neg P$, then Gödel's theorem tells us that PM is ω-inconsistent. And then there will be no way of defining in the model-theorist's sense (see footnote 14) the set of natural numbers *in* PM.

Putnam has also observed that our interpretation of the first part of the notorious paragraph may be naturally extended to account for Wittgenstein's final, puzzling remark that "what is called 'losing' in chess may constitute winning in another game."[18] Many readers (Bays and Steiner included) have seen in this remark a rejection of the idea of true but unprovable propositions. But on our interpretation Wittgenstein's remark makes much better sense. For if P were "false" in what Wittgenstein calls "the Russell sense"—that is, disprovable in PM (in which case, as we have said, PM would be, if consistent, ω-inconsistent)—if we imagine another system of set theory, S, which is strong enough to prove the consistency of PM, then *P will be a theorem of S*! This is a way of showing that what is called "losing" in one game may be called "winning" in another.

---

[15] See Hao Wang, *Reflections on Kurt Gödel* (Cambridge: MIT, 1987) pp. 84ff.; cf. J.W. Dawson, Jr., *Logical Dilemmas: The Life and Work of Kurt Gödel* (Wellesley, MA: A.K. Peters, 1997), chapter IV.

[16] *Kurt Gödel: Collected Works, Volume V, Correspondence H–Z,* Feferman et al., eds. (New York: Oxford, 2003), p. 133.

[17] The assumption that arithmetic is bivalent can itself be given two very different interpretations: it can mean either (1) that the schema $p \vee \neg p$ is accepted as part of the deductive apparatus of arithmetic, or (2) that every sentence of arithmetic is (in language introduced by Dummett, who repeatedly emphasizes the difference between these assumptions) that every sentence of arithmetic is "determinately true or false" (see Michael Dummett, *The Logical Basis of Metaphysics* (Cambridge: Harvard, 1991), pp. 74ff. Here we have in mind the stronger assumption.

[18] Putnam, "A Note on Steiner on Wittgenstein, Gödel, and Tarski."

Formally put, the argument is very brief:

Assume (1) PM $\vdash \neg P$, and PM is consistent.

Assume an S such that (2) $S \vdash$ Con (PM).

Then by (1), (2), and Gödel,

(3) $S \vdash \neg$ (PM $\vdash P$).

If we assume S uses the same definitions of the natural numbers, primitive recursive predicates, and so on, as PM, then (3) implies that

(4) $S \vdash P$.

Here is a sense in which, to repeat, "losing" in PM implies "winning" in S.

In closing, we must consider Bays's reasoning about the so-called "standard" model of arithmetic and the notion of *truth*. Throughout his paper Bays assumes that the notions of *truth*, *interpretation*, and *natural number* are and ought to be determined from within model theory alone, with reference—indeed, what Bays calls "perfectly canonical" reference—to the set (or model) $\mathbb{N}$ (210).

Bays never defines $\mathbb{N}$, but let us do so now. $\mathbb{N}$ is a *structure* in the model-theorist's sense, that is, a set conceived of as coming to us equipped with two distinguished elements, zero and one, and the functions of addition and multiplication. $\mathbb{N}$ is to be regarded as an interpretation for a corresponding first-order language L containing the usual apparatus of quantification theory, two binary function symbols, and two constant symbols. The *full theory of* $\mathbb{N}$, Th($\mathbb{N}$), sometimes called by model theorists "true arithmetic," is defined as the set of sentences of L that are satisfied in $\mathbb{N}$ in the sense of Tarski; Th($\mathbb{N}$) is thus defined recursively via Tarski's notion of *satisfaction in a structure*.

We are given by this construction that every sentence of L will either be in Th($\mathbb{N}$) or it will not be. What this definition does not, however, tell us is how to determine, for any arbitrarily given particular sentence of L, whether that sentence is or is not an element of Th($\mathbb{N}$). Indeed, by establishing that Th($\mathbb{N}$) is in principle not recursively axiomatizable, Gödel showed there *is* no way of effectively determining this. He thereby showed that Th($\mathbb{N}$) can have no categorical axiomatization—that, in other words, it must have nonisomorphic ("nonstandard") models.[19]

---

[19] Note that Gödel's proof applies to *any* first-order formalized language L^* sufficiently rich to express the usual arithmetical operations of addition, multiplication, and quantification on the natural numbers. Thus even if it can express more than the

Bays tries to eliminate the issue of ω-consistency by reiterating several times that ℕ is a "core" or "perfectly canonical" "interpretation" of "mathematical truth" applicable to every formula of PM and our natural language in the same way, across a wide variety of counterfactual contexts (cf. 204, 206, 208, 210). But why Bays thinks that we should assume that the notion of *interpretation* is reducible to the model-theorist's notion of ℕ he does not say. Nor does he explain what he means by a "perfectly canonical" "interpretation."

We for our part were careful to distinguish, as most philosophers do, between the notions of *interpretation, model*, and *translation*. In terms of the notorious paragraphs under discussion, the relevant contrast is between *Übersetzung* (translation) and *Deutung* (interpretation), a distinction which Wittgenstein's German and our discussion of the notorious paragraph respect. Unfortunately Bays slides, equivocally, among them.[20] On our view, whether a given sentence of PM does or does not have a model of a particular kind, subject to a particular assumption about PM's being ω-consistent, affects but does not exhaust the issue of whether, why, and to what extent we can translate or interpret in an informal sense sentences of PM into English apart from that assumption. As we explicitly wrote in our "Note," a kind of translation (a "correlation" or "transformation") of sentences of PM into English might be possible even if it turned out that there were no models, hence no admissible interpretations, of PM itself as a whole (FP 626). In short, the relations among the notions of *model, interpretation*, and *translation* are as complex as the relations of any of these notions to that of *meaning*: to assess them, we need to be clear about the context within which we are using them. Wittgenstein's insight, as we understand it, is but a special case of this more general point.

---

minimal language *L*, any such *L** will fail to recursively axiomatize a full theory of the structure consisting of the natural numbers and the functions of addition and multiplication on these numbers.

[20] Note, for example, Bays on p. 204 n.17: "the insistence that we limit ourselves to models of PA when we interpret arithmetic runs rather deep in Floyd and Putnam's paper. At one point they even suggest that, were we to find PA inconsistent, we should conclude that there are no admissible interpretations of arithmetic (see [FP] 626)." As is easily verified, we never spoke about all admissible interpretations of arithmetic (nor even about PA), but, following Wittgenstein, only about "admissible interpretations of PM" (Bays has simply misread and misquoted the pronoun "it" in the passage he cites). A similar running together of the notions of "model" and "interpretation" occurs when Bays mistakenly says that the "core" of our argument is the idea that "*all* 'admissible interpretations' [of arithmetic] take place on nonstandard models [because]...what a formula 'expresses' depends on the model at which we interpret it" (202–03).

To summarize: Bays assumes throughout his paper that his remarks have a direct bearing on philosophical problems about the notion of *truth*. Yet we find it odd that Bays finds it reasonable to grant (1) that there exist philosophical problems about arithmetical truth, and also (2) that the existence of model theory will resolve these problems. In fact, Bays seems to think that an appeal to $\mathbb{N}$ proves, not only that the Gödel sentence has a truth value, but that every sentence of arithmetic has a truth value—that, in other words, the principle of bivalence holds in arithmetic.

It is not that the authors of the "Note" are defenders of any philosophy on which bivalence fails in arithmetic.[21] But neither would we want to read intuitionists (Brouwer, Heyting) or nominalists, finitists, and so on out of the camp of philosophy. And we certainly would not claim that Brouwer, Goodman, and so on have been *refuted* by Tarski, Abraham Robinson, and the other founders of model theory.

Does Bays think otherwise?

JULIET FLOYD

Boston University

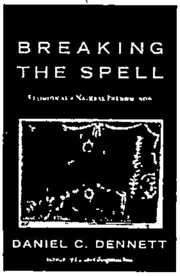
HILARY PUTNAM

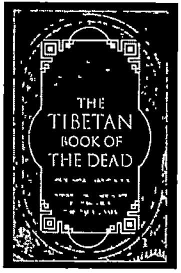
Harvard University

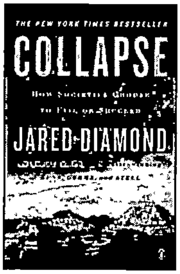
---

[21] See Putnam's Tarski lectures, "Paradox Revisited I: Truth," and "Paradox Revisited II: Sets—A Case of All or None?" in Gila Sher and Richard Tieszen, eds., *Between Logic and Intuition: Essays in Honor of Charles Parsons* (New York: Cambridge, 2000), pp. 3–26.

New titles from PENGUIN GROUP (USA)







DANIEL C. DENNETT
BREAKING THE SPELL
Religion as a Natural Phenomenon
Viking 448 pp. 0-670-03472-X $25.95

VOLTAIRE
CANDIDE or Optimism
Newly translated by Theo Cuffe
Introduction by Michael Wood
Penguin Classics 160 pp.
0-14-303942-3 $11.00

THE TIBETAN BOOK OF THE DEAD
Translated by Gyurme Dorje
Edited by Graham Coleman with Thupten Jinpa
Introduction by His Holiness The Dalai Lama
Viking 496 pp. 0-670-85886-2 $29.95

SAINT AUGUSTINE
CONFESSIONS
Translated by Garry Wills
Penguin Classics 304 pp.
0-14-303951-2 $16.00

JOHANN WOLFGANG VON GOETHE
FAUST, PART I
Translated with an Introduction and Notes by David Constantine
Preface by A. S. Byatt
Penguin Classics 352 pp.
0-14-044901-9 $10.00

JARED DIAMOND
COLLAPSE
How Societies Choose to Fail or Succeed
Penguin 592 pp. 0-14-303655-6 $17.00

G. I. GURDJIEFF
BEELZEBUB'S TALES TO HIS GRANDSON
All and Everything
FIRST SERIES
Tarcher 1,152 pp. 1-58542-457-9 $65.00

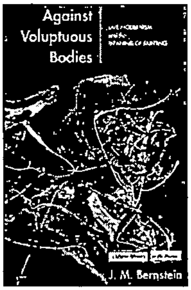
THE PENGUIN DICTIONARY OF PHILOSOPHY
SECOND EDITION
Edited by Thomas Mautner
Penguin 688 pp. 0-14-101840-2 $17.00

PLATO
EARLY SOCRATIC DIALOGUES
Edited with an Introduction by Trevor J. Saunders
Preface by Chris Emlyn-Jones
Penguin Classics 400 pp.
0-14-045503-5 $14.00

PHAEDRUS
Translated with an Introduction and Notes by Christopher Rowe
Penguin Classics 176 pp.
0-14-044974-4 $10.00

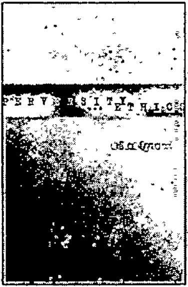
PENGUIN GROUP (USA) **www.penguin.com/academic**
Academic Marketing Department, 375 Hudson St., New York, NY 10014

NEW FROM
STANFORD UNIVERSITY PRESS



Against Voluptuous Bodies
Late Modernism and the Meaning of Painting

J. M. BERNSTEIN
The aim of this book is to provide an account of modernist painting—from Pollock to Ryman—that follows on from the aesthetic theory of Theodor W. Adorno. It offers a materialist account of modernism with detailed discussions of modern aesthetics from Kant to Arthur Danto, Stanley Cavell, and Adorno.
Cultural Memory in the Present
$29.95 paper $75.00 cloth

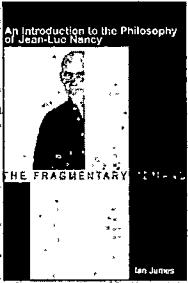


Perversity and Ethics

WILLIAM EGGINTON
Perversity and Ethics argues that a psychoanalytic reading of the phenomenon of perversity is crucial to understanding contemporary philosophical ethics.
$24 95 paper $60 00 cloth

The Fragmentary Demand
An Introduction to the Philosophy of Jean-Luc Nancy

IAN JAMES
This introduction to the philosophy of Jean-Luc Nancy gives an overview of his philosophical thought to date and situates it within the broader context of contemporary French and European thinking.
$19.95 paper $50 00 cloth



The Unthought Debt
Heidegger and the Hebraic Heritage

MARLÈNE ZARADER
Translated by Bettina Bergo
Drawing on Heidegger's corpus, the work of historians and biblical specialists, and contemporary philosophers like Levinas and Derrida, Zarader brings to light the evolution of an *impensé*—or unthought thought—that bespeaks a complex debt at the core of Heidegger's hermeneutic ontology.
Cultural Memory in the Present
$24.95 paper $65.00 cloth

Stanford
800.621.2736 www.sup.org University Press

Young Epistemologist Prize

The **Rutgers Epistemology Conference** announces the fifth bi-annual prize for the best essay submitted to the Conference by a person with a Ph.D. obtained by the time of submission but not earlier than ten (10) years prior to the date of the conference. Thus, for the Conference on May 4th/5th, 2007, the Ph.D. must not have been awarded prior to May, 1997. The prize winning essay will be presented at the bi-annual conference and will be published (along with the invited papers) in Philosophy and Phenomenological Research. The winner of the prize will receive an award of $1,000 plus all travel and lodging expenses connected with the Conference.

Details

The essay may be in any area of epistemology. The essay ***must*** be limited to 5,000 words (not including footnotes and bibliography) and submitted on a PC-formatted disk along with three (3) printed copies. The essays will be judged blindly and, thus, the three printed copies (but not the disk copy) must be prepared in such a way that the author's identity is completely masked. ***Only*** papers submitted in that fashion will be considered. Papers ***must*** be postmarked by November 25, 2006. The winner of the prize will be announced by February 10, 2007. Please include your email address in the cover letter accompanying your paper. By submitting the essay, the author agrees to have it posted on the Rutgers Epistemology Conference Webpage and published in Philosophy and Phenomenological Research.

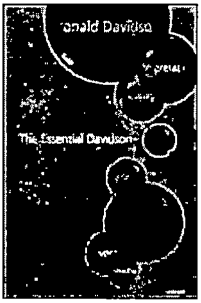
Please send all papers (along with the disk and a self-addressed postcard that will be returned to entrants acknowledging receipt of the essay) to:

The Young Epistemologist Prize
Philosophy Department
Rutgers University
26 Nichol Ave.
New Brunswick, NJ 08901-2882

For more information about the Rutgers Epistemology Conference and the Young Epistemologist Prize, please refer to:

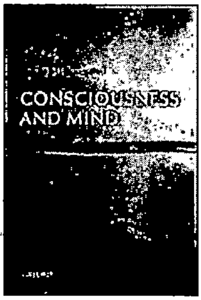
http://epistemology.rutgers.edu

New from
Oxford

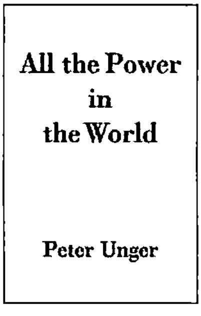


Donald Davidson

The Essential Davidson



CONSCIOUSNESS AND MIND



All the Power in the World

Peter Unger

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS

The Essential Davidson

DONALD DAVIDSON

The Essential Davidson compiles the most celebrated papers of one of the twentieth century's greatest philosophers. With a new introduction by Ernie Lepore and Kirk Ludwig, two of the world's leading authorities on his work, the volume distills Donald Davidson's seminal contributions in mind, action, language, and knowledge, from over three decades' worth of essays, into one thematically organized collection.

January 2006 paper $24.95 cloth $65 00

Consciousness and Mind

DAVID ROSENTHAL

David Rosenthal is one of the leading contributors to the philosophical study of consciousness. This volume gathers together his work on the subject from the past two decades, and represents the definitive presentation of his influential theory of consciousness as higher-order thought.

2006 paper $35.00 cloth $99.00

All the Power in the World

PETER UNGER

Peter Unger provocatively breaks with what he terms the conservatism of present-day philosophy, and returns to central themes from Descartes, Locke, Berkeley, Hume, and Russell. Wiping the slate clean, Unger works from the ground up to formulate a new metaphysic capable of accommodating our distinctly human perspective.

2005 $74 00

The Epistemology of A Priori Knowledge

the late TAMARA HOROWITZ

Edited by JOSEPH L. CAMP, JR.

This volume collects four published articles by the late Tamara Horowitz and two unpublished papers on decision theory: "Making Rational Decisions When Preferences Cycle" and the monograph-length "The Backtracking Fallacy."

2005 $45.00

Free Will and Luck

ALFRED R. MELE

Mele's ultimate purpose in this book is to help readers think more clearly about free will. He identifies and makes vivid the most important conceptual obstacles to justified belief in the existence of free will and meets them head on.

February 2006 $49.95

CAMBRIDGE

OUTSTANDING SCHOLARSHIP

Descartes
A Biography
Desmond Clarke
$40.00: Hardback: 0-521-82301-3: 520pp

The Spiritual Dimension
Religion, Philosophy and Human Value
John Cottingham
$70.00: Hardback: 0-521-84377-4: 198pp
$24.99: Paperback: 0-521-60497-4

Genetics and Christian Ethics
Celia Deane-Drummond
New Studies in Christian Ethics 25
$75.00: Hardback: 0-521-82943-7: 306pp
$29.99: Paperback: 0-521-53637-5

The Cambridge Companion to Kant and Modern Philosophy
Edited by Paul Guyer
Cambridge Companions to Philosophy
$75.00: Hardback: 0-521-82303-X: 736pp
$29.99: Paperback: 0-521-52995-6

The Cambridge History of Eighteenth-Century Philosophy
Edited by Knud Haakonssen
$140.00*: 2-Volume Hardback Set:
0-521-41854-2: c.968pp

From Apology to Utopia
The Structure of International Legal Argument
Martti Koskenniemi
$110.00: Hardback: 0-521-83806-1: 702pp
$50.00: Paperback: 0-521-54696-6

Linguistic Turns in Modern Philosophy
Michael Losonsky
The Evolution of Modern Philosophy
$70.00: Hardback: 0-521-65256-1: 296pp
$24.99: Paperback: 0-521-65470-X

Illegal Beings
Human Clones and the Law
Kerry Lynn Macintosh
$28.00: Hardback: 0-521-85328-1: 286pp

Lawyering Skills and the Legal Process
Caroline Maughan and Julian Webb
Law in Context
$45.00: Paperback: 0-521-61950-5: 468pp

Leo Strauss and the Theologico-Political Problem
Heinrich Meier
Modern European Philosophy
$60.00: Hardback: 0-521-85647-7: 208pp

Aristotle's *Nicomachean Ethics*
An Introduction
Michael Pakaluk
Cambridge Introductions to Key Philosophical Texts
$70.00: Hardback: 0-521-81742-0: 358pp
$27.99: Paperback: 0-521-52068-1

Hardwired Behavior
What Neuroscience Reveals about Morality
Laurence Tancredi
$28.99: Hardback: 0-521-86001-6: 240pp

The Judicial Process
Realism, Pragmatism, Practical Reasoning and Principles
E.W. Thomas
$95.00: Hardback: 0-521-85566-7: 442pp

Phenomenology, Logic, and the Philosophy of Mathematics
Richard Tieszen
$70.00: Hardback: 0-521-83782-0: 368pp

*Prices subject to change.

www.cambridge.org/us/philosophy/

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

NEW AND NOTEWORTHY
FROM CAMBRIDGE

The Grammar of Society
The Nature and Dynamics of Social Norms
Cristina Bicchieri
$70.00: Hardback: 0-521-57372-6: 274pp
$24.99: Paperback: 0-521-57490-0

Language and Mind
Noam Chomsky
$70.00. Hardback: 0-521-85819-4: 208pp
$24.99: Paperback: 0-521-67493-X

Aristotle and the Science of Nature
Unity without Uniformity
Andrea Falcon
$75.00: Hardback: 0-521-85439-3: 158pp

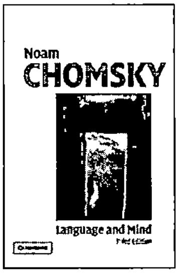
The Ascetic Self
Subjectivity, Memory and Tradition
Gavin Flood
$75.00: Hardback: 0-521-84338-3: 302pp
$29.99: Paperback: 0-521-60401-X

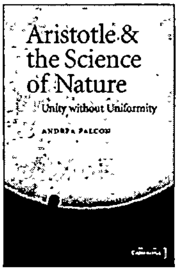
Kant's Cosmopolitan Theory of Law and Peace
Otfried Hoeffe
Modern European Philosophy
$70.00: Hardback: 0-521-82676-4: 272pp
$24.99: Paperback: 0-521-53408-9

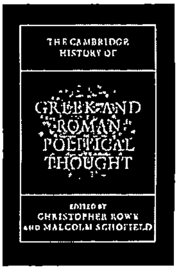
Sanctified Violence in Homeric Society
Oath-Making Rituals in the Iliad
Margo Kitts
$75.00: Hardback: 0-521-85529-2: 258pp

Theories of Judgment
Psychology, Logic, Phenomenology
Wayne Martin
Modern European Philosophy
$75.00: Hardback: 0-521-84043-0: 202pp

Now in Paperback...
The Cambridge History of Greek and Roman Political Thought
Edited by Christopher Rowe and Malcolm Schofield
with Simon Harrison and Melissa Lane
$48.00: Paperback: 0-521-61669-7: 766pp







Prices subject to change

www.cambridge.org/us/philosophy/

CAMBRIDGE
UNIVERSITY PRESS

Batta
30.8.06

THE JOURNAL OF PHILOSOPHY

VOLUME CIII, NUMBER 3
MARCH 2006

page
111 *The Invariance of Sense* Robert May
COMMENTS AND CRITICISM
145 *A New Problem of Descriptive Power* Stephan Leuenberger

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Christopher Peacocke, Carol Rovane, Achille C. Varzi. *Consulting Editors*: David Albert, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2006

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2006 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIII, NO. 3, MARCH 2006

THE INVARIANCE OF SENSE

How many senses can a given name have, with its reference held fixed? One, more than one? One answer that most would agree to is that sense is unique for each utterance of a name, that is, that a name can have no more than one sense on any given occasion. But is sense unique in any stronger sense than this? The answer that is typically attributed to Frege is that there is not, that, as Tyler Burge puts it, Frege "treats proper names as having different senses while applying to the same person."[1] There are a number of possibilities for the locus of this multiplicity of sense; the following remark by Ruth Barcan Marcus indicates the possibilities: "the sense of a term is whatever is grasped or understood by a speaker on a particular occasion of use and may vary from occasion to occasion as well as from speaker to speaker."[2] Of the views canvassed by Marcus, we can draw out a more conservative one, and one more extreme. On the more conservative view, Frege is holding that sense may vary from speaker to speaker; on the more extreme view, Frege holds not only this, but that sense may vary from context to context. Endorsements of the two views are not hard to find. For example, typical sorts of endorsement of the conservative view are found in Harold Noonan's remark that "different senses [are] associated with the name 'Aristotle' by different users,"[3] and Graeme Forbes's comment "that the sense of a proper name can vary from person to person."[4]

---

[1] Burge, "Sinning against Frege," *The Philosophical Review*, LXXXVIII (1979): 398–432, at p. 402.

[2] Marcus, "Review of Leonard Linsky: *Names and Descriptions*," *The Philosophical Review*, LXXXVII (1978): 497–504, at p. 503. This is not an interpretation of Frege that Marcus herself endorses; see her remark cited below.

[3] Noonan, *Frege: A Critical Introduction* (Cambridge, UK: Polity, 2001), p. 179.

[4] Forbes, "Indexicals and Intensionality: A Fregean Perspective," *The Philosophical Review*, XCVI (1987): 3–31, at p. 5.

0022-362X/06/0303/111–44

John Perry chimes in with the claim that "In the case of proper names, Frege supposes that different persons attach different senses to the same proper name."[5] Saul Kripke is another booster; in his view, Frege "concluded that, strictly speaking, different speakers of English (or German) ordinarily use a name such as 'Aristotle' in different senses (though with the same reference)."[6] Advocacy of the more extreme view is to be found in Burge's assertion that "Frege makes...remarks emphasizing the variability of the sense of a proper name for different users and in different contexts" (*op. cit.*, p. 403). On such views, even though sense is not shared, communicability between speakers need not suffer so long as speakers merely grasp that the senses they grasp determine the same reference; given the sense/reference distinction, the onus for accounting for the efficacy of language as a medium of communication shifts from shared sense to common reference. Let us call the sort of view being described the *variability view* of sense.

It seems to me, common wisdom notwithstanding, that nearly as one can make out, the variability view of sense is unquestionably incorrect as a reading of Frege. Frege's view, I believe, is quite clear: a proper name has a unique sense, not only *qua* utterance, but also in that sense cannot vary from speaker to speaker, from utterance to utterance, nor from context to context, and no speaker can associate more than one sense with a given proper name. That is, Frege held an *invariance* view of sense. To be somewhat more precise, I will argue that Frege's view is that *if* speakers speak the same language, then they associate the same sense with any given proper name; thus, if they did not associate the same sense, they would not speak the same language. Indeed, if sense varied from context to context, or even worse, from utterance to utterance, then there would be no stable notion of shared language at all, or at least not one that would be sustained by any notion of speakers sharing a common stock of thoughts, and this would cut deep for Frege, given his conception of language, logic, and thought. As Marcus observes, following up on the comment above, "such a view, if taken as...the only account of meaning in Frege, would be indefensible" (*op. cit.*, p. 503).

---

[5] Perry, "Frege on Demonstratives," *The Philosophical Review*, LXXXVI (1977): 474–97, at p. 485.

[6] Kripke, "A Puzzle about Belief," in Avishai Margalit, ed., *Meaning and Use* (Boston: Reidel, 1979), pp. 239–83 at p. 240. Kripke also attributes this view to Russell. This attribution is closer to correct; Russell held that different speakers may take a given name to abbreviate different descriptions. In "Knowledge by Acquaintance and Knowledge by Description" (in Russell, *Mysticism and Logic* (New York: Norton, 1929), pp. 209–32), he remarks that "the thought in the mind of a person using a proper name can generally only be expressed explicitly if we replace the proper name by a description. Moreover, the description required to express the thought will vary for different people, or for the same person at different times. The only thing constant (so long as the name is rightly used) is the object to which the name applies. But so long as this remains constant, the particular description involved usually makes no difference to the truth or falsehood of the proposition in which the name appears" (p. 216). Russell's view is rooted in his epistemology; the description by which a speaker abbreviates a proper name characterizes knowledge the speaker has of the object denoted. Part of my argument to follow is that it is incorrect to assimilate Frege's views to Russell's, in part because Frege's view of the relation of sense and knowledge is not directly comparable to Russell's notion of knowledge by description.

My object in this paper will be to show that Frege holds that the sense of a proper name is both unique and invariant, and that the sense of a name does not vary in any way. This is fundamentally a linguistic fact for Frege; it is not merely a reflection of speakers' deference to shared social or communicative practices. It is not that sense is constant from speaker to speaker by convention.[7] But if Frege's view is conditionalized to language, the notion of language appealed to is a peculiarly Fregean one, and so we will first examine Frege's view of language. We will then consider the passages from Frege's writings that have largely animated the discussion, drawn from "The Thought," and from "On Sense and Reference." Both of these, in my view, have been subject to rather strained interpretations, and have glossed over important comments by Frege. In the passage from "The Thought," we will see that Frege clearly enunciates the invariance view, and we will see how this view is compatible with remarks from the famous "Aristotle" footnote in "On Sense and Reference" that have been taken as Frege's endorsement of the variability view. From this will emerge two issues—the relation of sense and modes of presentation to descriptive conditions, and Frege's views of the cognitive relation in which speakers may stand to senses.

Lest one think that the issue of concern here is merely one of Fregean exegesis, we note that the variability view has been taken, especially in its more radical form by Burge and Perry, to underlie far-reaching claims about the significance of the doctrine of sense and reference. This is in large part because it appears to be at odds with Frege's view of the universality of meaning. If meaning is something that is shared by speakers of a language, and so invariant from speaker to speaker, then if sense is not necessarily constant from speaker to speaker, or even utterance to utterance, it would appear that sense is not a suitable candidate for the meaning of a proper name. But if this holds for proper names, then, given Frege's view of the ubiquity of naming, in general sense cannot be identified with meaning. A main consequence of what I will be arguing is that this dramatic conclusion is unwarranted, given its dependence on a view that Frege does not hold about how many senses a proper name can have.[8]

---

[7] This view is most commonly associated with Michael Dummett; see chapter five of *Frege: Philosophy of Language* (London: Duckworth, 1973); as he remarks there "The notion of sense is thus of importance, no so much in giving an account of our linguistic practice, but as a means of systematizing it" (p. 105).

I. FREGEAN LANGUAGES

As standardly conceived, a language is a set of *sentences*, where sentences are concatenated strings of *symbols*. As such, languages are characterized in a purely formal manner, sentences and their constituent symbols being formal marks. Languages are defined through a rule system that specifies the well-formed concatenations of symbols, the standard procedure being to categorize the symbols (the vocabulary of the language), and then state the rules in terms of these categories. Languages so conceived are the same if and only if they are the same set of sentences; otherwise, languages are distinct, even if they overlap in membership.[9] Thus, languages can differ either because they contain different symbols, or, with the same symbols, because the symbols are concatenated in different ways. Understanding language in this way has been particularly important in those areas of inquiry, in linguistics, logic, philosophy, mathematics, and computer science, where computation is of central concern. In all of these areas computation is understood, roughly, as symbol matching; a computational procedure mapping strings of symbols onto strings of symbols can be executed insofar as some formal criterion is satisfied. We have accrued substantial knowledge of language so conceived, including, especially since the seminal work of Noam Chomsky,[10] of its combinatoric and computational properties.

---

[8] In this paper, I focus exclusively on the senses expressed by proper names. In the sequel, "Frege on Indexicals" (forthcoming in *The Philosophical Review*), I discuss this issue as it arises with indexicals and demonstratives. Reflection on the putative sense variance of these latter expressions has led to the development of replacement notions argued to capture the insights for which Frege postulated senses, but which satisfy criteria for having the status of linguistic meanings; we have in mind here Perry's notion of *role* and David Kaplan's concept of *character*, see Perry, and see Kaplan, "Demonstratives," in Joseph Almog, Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford, 1989), pp. 481–563.

[9] It is standard to distinguish between languages, *qua* sets of sentences, being weakly and strongly equivalent, that is, whether they are the same set of concatenated strings of symbols, or are the same set of structural descriptions of those strings. Although the term *sentence* is commonly used to cover either of these notions, for our purposes here we can gloss over this distinction.

[10] I have in mind here the Chomsky-hierarchy that categorizes languages in terms of their (weak) generative capacity. See Noam Chomsky, "Formal Properties of Grammars," in R.D. Luce, R.R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology*, Volume II (New York: Wiley, 1963), pp. 323–418.

It is a fundamental attribute of languages as sets of sentences that they are inherently *uninterpreted* systems; that is, whatever properties such systems have, they are strictly formal properties, and languages are individuated in terms of these formal properties. However, such languages may be assigned an interpretation, again by standardly known procedures: a model, including a stipulation of a universe of values, is specified, and an interpretation is defined on that model that maps from (nonlogical) expressions of the language to values in the model. On this method, the interpretation of symbols of the language is not fixed, but may vary, given the model; the language itself, however, remains unchanged under this variance of interpretation. This model-theoretic approach to interpretation has proved to be extraordinarily fruitful, especially in the characterization of metalogical properties of languages used as logistic systems. To take one example from an extensive catalogue, one of the earliest illustrations of this method was by Hilbert, who showed the independence of the parallel postulate from the other axioms of Euclidian geometry by providing a model in which the former is false and the latter are true.[11]

While there are aspects of this view of language and interpretation that arguably originate with Frege (for example, the notion of computation described above derives from Frege's view of the proper conduct of proof), I have gone to pains to describe it in order to place it in relief to Frege's very different perspective, since as a whole this is *not* the way that Frege conceives of a language.[12]

For Frege, a language is not a system of marks or symbols (individuated solely by their shapes, what Frege calls a *figure*[13]), but rather as a system of *signs.* A sign is a pairing of a symbol and a sense, and as such it can be used, unlike a symbol *simpliciter*, for the purpose of reference. Concatenations of signs labeled by Frege *propositions* are themselves complex signs, one component of which is a thought, itself a concatenation of senses. A language for Frege is thus a set of propositions, not sentences; as such, it is a set of correlations between concrete things—symbols—and abstract things—senses. Knowing a language is thus knowing such a system of correlations, where we come to know the senses via our grasp of them.

---

[11] Note that we can investigate the semantic structure of a given language with respect to a fixed model, that is, one with a specified interpretation relative to a given universe; this is what is standardly done in model-theoretic explorations into the semantics of natural languages. For current purposes bear in mind however that the notion of an interpreted language that emerges in this manner is distinct from that we find with Frege's conception of language that we discuss below.

[12] The remarks in the following paragraphs develop the discussion in Aldo Antonelli and Robert May, "Frege's New Science," *Notre Dame Journal of Formal Logic*, XLI (2000): 242–70, in section 2.

[13] In his critique of formalism found in "On Formal Theories of Arithmetic" (1885), translated by Eike-Henner W. Kluge in Kluge, ed., *On the Foundations of Geometry and Formal Theories of Arithmetic* (New Haven: Yale, 1971), pp. 141–53, especially in the discussion on pp. 114–16.

Because a Fregean language is composed of signs, observe that it follows that it is an *interpreted* system. Sense uniquely determines reference, so that difference of reference entails difference of sense; consequently, it is impossible in a *given* language, for a sign to have a reference different than what it has. This holds for all expressions of a language, including propositions. Their reference, to the True or the False, is determined by their thought components, and so each proposition comes immutably equipped with one and only one truth value. It makes no sense, on Frege's view, to consider a true thought (that is, a thought whose reference is the True) to be false, for to do so would not be to consider that thought, but another. Thus, it is a consequence of the doctrine of sense and reference that a language is an inherently and uniquely interpreted system. To Frege, it would be nonsensical to speak of an interpreted language in the sense of an interpretation being *assigned* to a language, for this presupposes something that Frege rejects, that a language itself is a system of meaningless marks or symbols.[14] To emphasize again the point here: a change in the system of signs is not for Frege just a different interpretation of the language, but gives a different *language* altogether.

For Fregean languages, the criteria for the individuation of languages turns on whether there is the same set of propositions. Thus, *qua* language as a system of signs, any alteration to the pairing of symbols and senses would be an alteration of the signs, and that would constitute a different system of signs, and so a different language altogether.[15] If there is a change of signs, then there is a change of language. Given that a sign is a complex, signs can differ because one or the other of their coordinates differ; the case that interests us is where the symbols remain the same yet there is a difference of signs because the senses with which those symbols are paired are different. Thus, we could have two Fregean languages that are the same with respect to strings of symbols, but yet are different languages because those symbols are associated with different senses. Such languages would differ as sets of propositions. Note that this would be so even if the *references* of the symbols in the two languages were exactly the same; whether reference varies with difference of sense is not material to whether we have distinct languages or not.

---

[14] Put in contemporary parlance, Fregean languages do not contain any nonlogical constants; there can be no schematization in Fregean languages, universality can only be expressed through generalization. This point is due to William Demopoulos in his "Frege, Hilbert, and the Conceptual Structure of Model Theory," *History and Philosophy of Logic*, xv (1994): 211–25; see also Antonelli and May, and the discussion in footnote 11.

[15] As discussed in Antonelli and May, while the symbols of a language are an arbitrary characteristic of a particular language, there is a universality inherent in language in that senses are equi-expressible with respect to any language whatsoever. Against this, the standard way that languages will differ is because they incorporate different sym-

Our description of Fregean languages as constituted as sets of propositions is meant to characterize language in its most general sense, so that falling under this characterization of language is any system, either natural or designed, used for the expression of thoughts. It thus includes English and German as well as the *begriffsschrift*, Frege's conceptual notation language. On Frege's conception, nested within the class of languages generally conceived there is a more circumscribed class of languages that Frege referred to as *logically perfect*. These are languages that given their design are properly fit to use for reasoning, that is, languages in which proofs can be rigorously carried out. For Frege, a proof is an ordered sequence of propositions; what a proof shows is how true thoughts follow from other true thoughts. Establishing that a sequence of propositions *is* a proof, however, is something that is determined strictly in virtue of the symbolic forms of the propositions; to specify the rules by which proofs may legitimately proceed from step to step, we need only take into account propositions *qua* sentences, that is, concatenations of symbols. Signs are logically manipulated in terms of their symbolic structure; one proposition follows from another because their *forms* stand in some specified relation.[16] The contents they express, which is what is proven, is not directly material to whether it can *be* proven.

Given this conception, we can be certain of the reliability of proofs to the extent that each sentential string of symbols expresses one and only one proposition; a language will qualify as logically perfect to the extent that it satisfies this condition. In the design of a language, this can be facilitated if all the signs are pair-wise disjoint from each other,

---

bols, although we can still have different languages even if the symbols and senses are invariant, if the mappings between the symbols and senses were different in the languages. In that case, we would have distinct sets of propositions.

[16] For Frege, this relation is governed by Modus Ponens, which is the only rule of inference, in that the transitions from propositions to propositions on the right-hand side of proofs in the *Begriffsschrift* notation are always Modus Ponens steps (or are reducible to Modus Ponens steps). For discussion of rules of inference in Frege's logic, see my notes "Notes on Frege on Rules of Inference," at http://kleene.ss.uci.edu/~rmay/Inference.pdf.

so as to insure that they differ from each other in both their symbol and sense coordinates. In that case, since any two distinct symbols (*qua* their formal shape) are distinct signs, we can be assured that each symbol—that part of the sign open to inspection—will express a unique sense. Not all languages measure up to this design standard, for they will have distinct signs that differ in only one of the coordinates. Natural languages are notorious offenders. Unlike logically perfect languages, such as the *begriffsschrift,* they will contain homonyms, signs whose symbols agree, but whose senses do not.[17]

Nested within the logically perfect languages, we must recognize yet one other sort of language that Frege isolates. These are logistic systems, languages in which reasoning, that is, proofs, are actually carried out. They are distinguished from a logically perfect language *simplicter* in that certain propositions formulable in the language are designated as basic, and other propositions are proven from them, by the specified rules of inference (always modus ponens), as theorems. Among the logistic systems, the narrowest is pure logic, in which we only require the logical vocabulary (the logical connectives, quantifiers, and variables of appropriate levels, and so on) to state the stipulated, unproven propositions, what Frege called the *Basic Laws.*[18] In addition, there are broader systems of applied logic, which in addition to the Basic Laws, contain other basic propositions that Frege labels *axioms.* These differ from Basic Laws in that their statement involves nonlogical vocabulary, so that they will have application specific content that is lacking from any Basic Law. So, consider geometrical axioms, whose statement will include (along with logical vocabulary), terms referring to points, lines, and so on.

In what follows, we will be considering a number of remarks by Frege about language, and when doing so it will be important, if we are to understand properly their import, to bear in mind the distinctions we have drawn out among languages, as Frege in his remarks often shifts among the various notions, sometimes intending his comments to apply to language in general, other times to some narrower conception. In general, Frege's custom is that when he makes a positive point with illustrations from natural language, he means to be saying something about language in general. A typical example would be Frege's discussion in "On Sense and Reference" of the sense/reference distinction, which is clearly meant to apply to all languages. When Frege wishes to make a point about language that applies with less generality, his custom is to make this explicit. Consider for instance Frege's discussion of the referential failure of proper names in natural language; propositions that contain them can only express "mock," not real, thoughts.[19] This is, Frege is explicit to point out, a linguistic *im*perfection, and not an aspect of a logically perfect language. On the other hand, when Frege is laying out the characteristics of a logically perfect language, for instance, in the opening sections of *Begriffsschrift* or *Grundgesetze*, he uses explicit formalism, and carefully explicates the intended meaning of the terms introduced. Without such provisos, however, Frege's intention is normally to speak without restriction about language, perfect or imperfect, even if his examples are drawn from languages that are within the latter class.

---

[17] Natural languages will also contain synonyms, signs which differ as symbols, but not as senses, which are also ruled out by the condition, and hence also will not occur in logically perfect languages. From the point of view of proof, this exclusion would amount to saying that a thought could be proven in the guise of one and only one proposition. In at least one sense, this is too strong, given Frege's understanding of explicit definition as identifying the sense of the defiendum with that of the defiens. Thus, our intention is to exclude from logically perfect languages only synonyms that are not definitionally related; this can be achieved by limiting the condition stated to just atomic signs.

[18] Frege held that one sort of object was nameable in terms of pure logic, namely, value ranges, the purely logical objects. This assumption was of course highly problematic, given the inconsistency of Basic Law V, the identity condition for value ranges.

Although there is much more to say about Frege's views, as well as their genesis, this completes our review of the Fregean conception of language, understood as a set of propositions, rather than sentences.[20] I am, however, now in a position to give a somewhat more precise rendering of the thesis that I will be arguing *against*: A symbol can change its associated sense, and still remain the same sign, presumably in virtue of the determined reference remaining the same. This is the thesis that underlies the variance view. Frege's explicit view, however, is to the contrary; his view is that any such change in sense would give rise to a different sign, and indeed, to a different language.

II. THOUGHTS AND PROOF

As came up in the discussion in the previous section, Frege's had a peculiar view of proof. For him, what is shown in a proof is how a given true thought follows from other true thoughts. While Frege is justly celebrated for his innovation of insisting on the rigor of proof, this was a claim about how proofs are properly conducted, and not about what is proven. We can only determine that a thought follows from other thoughts in virtue of formal aspects of the symbolic part of the signs by which the thoughts in question are expressed. Frege summarizes this view of proof in the following remark:

---

[19] See in particular the discussion in "On Sense and Reference" (1892), translated by Max Black, in Peter Geach and Black, eds., *Translations from the Philosophical Writings of Gottlob Frege* (Oxford: Blackwell, 1952), pp. 62–63; the terminology of "mock thoughts" is to be found in "Logic" (1897), translated by Peter Long and Roger White, in H. Hermes, F. Kambartel, and F. Kaulbach, eds., *Posthumous Writings* (Chicago: University Press, 1979), pp. 126–51, on p. 130. We turn to the significance of such terms in section VII.

[20] For discussion of the central role Frege's conception of language plays in his debate with Hilbert over independence proofs, see Antonelli and May, and the remarks in footnote 15.

> What can be interpreted is perhaps a symbol or a group of symbols, although the univocity of the symbols—which we must retain at all cost—excludes different interpretations. But an inference does not consist of symbols. We can only say that in the transition from one group of symbols to a new group of symbols, it may look now and then as though we are presented with an inference. An inference simply does not belong to the realm of symbols; rather, it is the pronouncement of a judgment made in accordance with logical laws on the basis of previously passed judgments. Each of the premises is a determinate thought recognized as true; and in the conclusion, too, a determinate thought is recognized as true. There is here no room for different interpretations.[21]

Now it should be clear what Frege is saying. What is proven is a *particular, determinate thought,* although this thought is accessible to us only through its symbolic form; that some set of transitions from propositions to propositions constitutes a proof of this thought depends upon this form. Now suppose, in accordance with the variance view of sense, that the sense associated with a sign could change. It might be thought that any such change would leave proofs unaffected, for there would be no change in the reference of any sign, and thus there could be no change in the truth value of any proposition. But to think that such change of sense would be innocent would be to misunderstand a central aspect of Frege's conception of logic. This is because if some constituent sense of a thought were to change, then what we would have would be a different thought altogether. But if a sentence that expresses a thought T can also express a distinct thought T^*, then if that sentence stands formally as the conclusion of a proof, how can we say whether it was the thought T or T^* that was proven? Proof would not be determinate; there would be no one definite *thought* that was proven. In such a language there could be no proofs in Frege's view; there would be no logic. Frege makes this point quite explicitly in the following passage:

---

[21] "On the Foundations of Geometry *II*" (1906), translated by Kluge in Kluge, ed., *On the Foundations of Geometry and Formal Theories of Arithmetic,* pp. 49–112, on p. 82. In this passage I have altered the translation of the German *zeichen,* which Kluge translates as "sign," to "symbol," so that it accords with the usage we have established in the previous section.

> As if it were permissible to have different propositions with the same wording! This contradicts the rule of unambiguousness, the most important rule that logic must impose on written or spoken language. If propositions having the same wording differ, they can do so only in their thought-content. Just how could there be a single proof of different thoughts? (*ibid.*, p. 79).

Note carefully the second sentence: "This contradicts the rule of unambiguousness, the most important rule that logic must impose on written or spoken language." Here Frege gives a ringing endorsement of the invariance view, for it is this view that accords with fundamental tenets of logical reasoning; to suppose that a symbol could be associated with more than one sense would violate these tenets.

In the continuation of the last remark, Frege provides a diagnosis of the source of the view he is repudiating:

> This looks as though what is proved is the wording alone, without the thought-content; and as though afterwards different thoughts were then supposed to be correlated with this wording in the different disciplines. Rubbish! A mere wording without a thought-content can never be proved (*ibid.*, p. 79–80).

The context of this remark is Frege's dispute with Hilbert as found in Frege's essays *Über der Grundlagen der Geometrie* (*op. cit.*). In it, Frege accuses Hilbert of a mistaken conception of logic, one that does not recognize that it is thoughts that are subject to proof. It is a conception of logic, according to Frege, that ultimately effaces the sense/reference distinction. It is also a conception of logic that has a distinctively contemporary ring; indeed, in this regard modern logic has followed Hilbert's, not Frege's, lead; a "mere wording" *can* be proved, with the words open to various interpretations. Perhaps the tendency to attribute the variance view of sense to Frege arises from an inclination to read aspects of the modern conception back into Frege's. But this would be a mistake, and a misreading of Frege's views.

III. DR. GUSTAV LAUBEN

In "The Thought," Frege makes the following remark:

> ...for every proper name there shall be just one associated manner of presentation of the object so designated.[22]

---

[22] "The Thought" (1918), translated by Geach and R.H. Stoothof (as "Thoughts") in Geach, ed., *Logical Investigations* (Oxford: Blackwell, 1977), pp. 1–30, on p. 12. The German is as follows:

Given, as Frege says in "On Sense and Reference," that the sense is wherein the manner of presentation is contained, this appears to be a clear endorsement of the invariance view.

This remark is drawn from Frege's well-known, if not notorious, discussion of Dr. Gustav Lauben and associates. It is of importance in this context because, Frege's pronouncements notwithstanding, it might be maintained that the argument in the previous section for the invariance view finds its application solely within logically perfect languages. In "The Thought," however, Frege couches the discussion in terms of natural languages, and so we can be more comfortable with the claim that it is meant to apply to languages in general. But before we delve into the text itself, we need to clarify somewhat more the identity criteria that apply to Fregean languages.

The criterion of identity for Fregean languages, as we observed in section I, is that languages are the same if and only if they are the same set of propositions, and that they can be the same set of propositions only if their vocabularies of signs are the same. Now suppose that we have the Fregean languages L_1 and L_2 such that the *symbols* of L_1 and L_2 are the same and they have the same syntax, so that as sets of *sentences* L_1 and L_2 are indistinguishable; that is, they are *formally identical* languages. Nevertheless, L_1 and L_2 are distinct as *Fregean* languages because, we assume, they differ in the pairing of symbols with senses; the signs they contain are not the same, and so L_1 and L_2 are not the same set of propositions. In this circumstance, since L_1 and L_2 are identical as sets of sentences, it is appropriate, for some sentence *S*, that we can speak of the *same* sentence in both languages, although not of the same proposition. In this case, we will say that L_1 and L_2 are *S*-equivalent, but not *P*-equivalent, languages.

Consider now a circumstance, unlike that of the previous section, in which we have two speakers, S_1 and S_2, such that S_1 speaks L_1 and S_2 speaks L_2. And, let us assume that L_1 and L_2 are, as described,

---

> Es muß also eigentlich gefordert werden, daß mit jedem Eigennamen eine einzige Weise verknüpft sei, wie der, die oder das durch ihn Bezeichnete gegeben sei. Daß diese Forderung erfüllt werde, ist oft unerheblich, aber nicht immer.

While in what follows translations are drawn from the Geach and Stoothof version, their translation of the ellipsis in the cited passage, which reads "So we must really stipulate that," does not capture as well Frege's intention as A. and M. Quinton's translation in *Mind*, LXV (1956): 289–311, at p. 298: "So it must really be demanded that...." The remainder of the Quinton's translation of this passage is however rather convoluted: "a single way in which whatever is referred to is presented be associated with every proper name."

S-equivalent, but not *P*-equivalent. Then, it follows that S_1 and S_2 do not speak the same language. They could only be said to speak the same language if they associate the same senses with the symbols that the languages they speak share; that is, if the languages they speak are *P*-equivalent. If S_1 and S_2 associate different senses with some symbol, then they do not speak the same language, even if the languages are *S*-equivalent.

Now suppose something further, that L_1 and L_2 have the additional characteristic that not only do they agree in their symbols, but that they also agree in their references; that is, each symbol-identical pairing of signs from L_1 and L_2 have the same reference, (although not the same sense). That is, L_1 and L_2 are not only *S*-equivalent but are also, in the sense just described, *R*-equivalent (but not *P*-equivalent.) Question: Can S_1 and S_2 speak intelligibly to each other? Can they communicate, even though strictly they speak different languages? The answer is that for practical purposes, yes. So long as speakers use the same symbols to refer to the same things, communication can proceed apace. It will not be of the moment how those references are presented to each of the speakers. All that is of relevance is what is being spoken about, that is, the reference, and not that speakers of L_1 and L_2 will have different beliefs about that reference.[23] Thus, all we require from this point of view for L_1 and L_2 to be identical is that they are *S*-equivalent and *R*-equivalent. This criterion will of course be satisfied if L_1 and L_2 are *P*-equivalent, since then they truly are one and the same language. But it may also be satisfied when they are not *P*-equivalent, in which case that they are in fact different languages may be effaced.

Notice that it may be opaque to S_1 and S_2 that they are speaking different languages. They may very well think they are speaking the same (Fregean) language, since given what is accessible to them—the symbols and perhaps their references—the languages they speak are nondistinct, since they are *S*- and *R*-equivalent. But then, given that they believe the languages they speak conform to principles of proper language design, in particular, the bi-unique relation of symbols and senses, they will think, in the absence of evidence to the contrary, that their languages are made up of the same signs. This assumption of identity of language, though false, thus will be innocent in most communicative contexts, so long as the references determined by the symbol-identical signs are the same.

---

[23] Because the customary sense is the indirect reference, if speakers associate different senses with a given sense, they will have different beliefs about that reference.

The view I have just presented is just, with some elaboration, that which Frege presents in the passage from "The Thought" from which the remark cited at the beginning of the section is drawn. The passage begins with Frege telling us that Herbert Garner and Leo Peter associate different senses with the proper name "Dr. Gustav Lauben." He argues to this conclusion by stipulating that Garner and Peter *know* different and disjoint things about Lauben; then if all information characteristic of the object "presented" attained by the grasp of sense is knowledge of that object, it follows that Garner and Peter must grasp different senses. The consequence of this is to divide linguistically Garner and Peter:

> Then as far as the proper name 'Dr. Gustav Lauben' is concerned, Herbert Garner and Leo Peter do not speak the same language, although they do in fact refer to the same man with this name; for they do not know they are doing so.[24]

What Garner and Peter do not know is that they are not speaking the same language, (they may know that they are referring to the same person). Frege continues:

> Therefore Herbert Garner does not associate the same thought with the sentence 'Dr. Gustav Lauben was wounded' as Leo Peter wants to express with it.

Indeed he could not, given the postulated difference in the languages they speak. Frege now suggests a workaround, in order to "avoid the awkwardness that Herbert Garner and Leo Peter do not speak the same language." After all, it appears to the untutored that Garner and Peter are both speaking German. This is to be accomplished by supposing:

> that Leo Peter uses the proper name 'Dr. Lauben' and Herbert Garner uses the proper name 'Gustav Lauben'.

The workaround brings the relevant signs in Garner's and Peter's languages into conformity with proper linguistic design principles by making them distinct not only as senses, but also as signs; they are now, we can say, *properly* distinct, and so both may be signs in a single language.[25] Now there is no bar to Garner and Peter speaking the same language, one which contains the names "Dr. Lauben" and

---

[24] This and the following passages are drawn from Frege, "The Thought," p. 12.

[25] Frege discussion of this case anticipates aspects of the Paderewski puzzle introduced by Kripke; see Robert Fiengo and May, *De Lingua Belief* (Cambridge: MIT, forthcoming), chapters 3 and 4, where the relation of these case is explicitly brought out.

"Gustav Lauben" (but not "Dr. Gustav Lauben"), each of which expresses a different sense.[26] Garner, Frege proceeds to tell us, may take:

> the sense of the sentence 'Dr. Lauben was wounded' to be true but is misled by false information into taking the sense of the sentence 'Gustav Lauben was wounded' to be false.

"So," Frege says, "given our assumptions these thoughts are different," since only if the thoughts expressed by the two sentences are different could Garner misjudge one, something he could not do if they expressed the same thought.

Frege has used the artifice of symbolic revision in order to make a point about distinctness of thoughts in languages, such as those we use in everyday life, that depart from logical perfection in that they are, as we have put it, only *S*- and *R*-equivalent. With respect to such identity of languages, with proper names, for every "way that the object so designated is presented...there corresponds a special sense of a sentence containing the proper name." All of the distinct thoughts into which these senses will enter will have the same truth value, since the languages in question are *R*-equivalent. But even so we must not lose sight of the fact that they are indeed different languages; we must not lose sight that "for every proper name there shall be just one associated manner of presentation of the object so designated." We should not, Frege tells us, lose sight of this, that is, the invariance thesis, even though for certain purposes of communication, "it is unimportant that this demand be fulfilled."[27]

To summarize, we can do little better than Frege's remark from "The Thought":

> To be sure the same words, on account of the variability of language with time, may take on another sense, express another thought; this change, however, relates only to the linguistic realm.[28]

---

[26] A subtlety here: To say that Garner and Peter speak the same language is not to say that what they grasp is the same, for it may be that Garner has the name "Gustav Lauben" in his language, but not "Dr. Lauben," while Peter conversely has "Dr. Lauben," but not "Gustav Lauben." This would be no bar, however, to saying that they partake of the same language.

[27] Burge, p. 403, says that "In 'The Thought' (pp. 516–518), Frege makes similar remarks emphasizing the variability of the sense of a proper name for different users and in different contexts." (The "similar" here refers to remarks from "On Sense and Reference" that we turn to in the next section.) The pages Burge cites (from the Quinton's translation), include the material discussed above; however, Burge does not develop any textual analysis of these passages in support of his claim about proper names; in particular, he does not discuss Frege's apparent direct assertion to the contrary. As it stands, Burge's claim is little more than an unsupported claim regarding Frege's views of proper names, one which there is very little reason, if any, to believe.

[28] Frege, "The Thought," p. 28.

To the extent that we find variance of sign, in that different senses are associated with a single symbol, it is because there are different languages. This variance can arise, as Frege points out in this remark, because of the diachronic change of language, but it can also obtain between speakers synchronically, as in the circumstances we have been discussing. But these are matters, as Frege rightly says, of language, of how languages are individuated. In any given language, however, there is one, and only one sense associated with each symbol; anyone who speaks this language grasps that association, and no other.[29]

IV. "ON SENSE AND REFERENCE," FOOTNOTE 2

One way of thinking about what Frege is trying to do in the discussion in "The Thought" is to demarcate a line between what constitutes knowing a language and what is needed to use a language. If we ask, according to Frege, whether Garner and Peter know and speak the same language, then the answer is that they do not. But for the purposes of communication, weaker criteria may be sufficient, and, as long as we are careful, we can say that in this realm, Garner and Peter do speak the same language. In "On Sense and Reference," in the famous second footnote, Frege makes much the same points, although there is a more subtle argument being presented because Frege here brings to the discussion an explicitly cognitive dimension as to what speakers *think* the sense of a proper name is (even when they speak the same language).

After opening "On Sense and Reference" with his discussion of identity statements, Frege spends the next number of paragraphs elucidating what he means by the sense of an expression. In the course of doing so, he says that "The sense of a proper name is grasped by everybody who is sufficiently familiar with the language or totality of designations to which it belongs" (*op. cit.*, pp. 57–58). Given his use of the definite description, implying that there is one and only one sense of a given proper name, Frege apparently felt that some clarification was needed at this point, which he provides in the following footnote:

---

[29] Frege makes much the same point in "On Sense and Reference." He says there that "To every expression belonging to a complete totality of signs, there should certainly correspond one and only one [*bestimmt*] sense; but natural languages often do not satisfy this condition, and one must be content if the same word has the same sense in the same context" (p. 58). The point Frege is making here is that in a properly designed language, there is sense invariance across contexts, and so such languages can be defined globally; in contrast, the definition of natural languages is inherently local, as they do not show this sort of stability, and in this regard depart from proper design. While it is not altogether clear what Frege means here by context, (historical linguistic epoch, as in the citation from "The Thought" in the text, or something more like context of utterance), we should nevertheless bear in mind the critical point emphasized in the text—natural languages do not depart from proper design in the sense that, *in any particular context, everyone who speaks a given language utilizes the same system of signs.* That is, within a context, for *any* language, sense invariance holds. Burge, for one, misses the point here; he observes that "The implication is clearly that the senses associated with proper names and other indexical constructions shifts with context" (p. 403), and takes this as an endorsement by Frege of the variance thesis. What he misses is that shift of sense with context in turn implies shift to a different language, and that for each of these languages, it is the invariance thesis that obtains.

> In the case of an actual proper name such as 'Aristotle' opinions as to the sense may differ. It might, for instance, be taken to be the following: the pupil of Plato and teacher of Alexander the Great. Anybody who does this will attach another sense to the sentence "Aristotle was born in Stagira" than will a man who takes as the sense of the name: the teacher of Alexander the Great who was born in Stagira. So long as the reference remains the same, such variations of sense may be tolerated, although they are to be avoided in the theoretical structure of a demonstrative science and ought not occur in a perfect language (*op. cit.*, p. 58).

What Frege is saying here seems little different than what he says in "The Thought." If our gaze is limited to *sentences*, and hence just to *S*-equivalent languages, then different speakers may associate different senses with the proper name "Aristotle" (although they in fact speak different languages), and this may be "tolerated" so long as all of these senses determine the same reference. But in circumstances in which what is paramount is not the exigencies of communication but rather the structure of the language, such as in a demonstrative science such as mathematics, we must meet a more stringent criterion of uniqueness of sense for proper names. We cannot, so to speak, play along as if there were just one language; rather we must insure that there *is* only one language in question, and that it meets the relevant standards of sense invariance.

Glossing the argument in this way presents it as a somewhat less articulated rendering of the argument presented in "The Thought." In particular, Frege does not draw out the consequence here that speakers who associate different senses with "Aristotle" speak different languages. There are, however, aspects of the argument that we find here that do not resonate in the presentation in "The Thought," and which perhaps give us a somewhat different and perhaps fuller sense of Frege's views on the senses of proper names. These devolve on Frege's comment that what differs in regard to the sense of "Aristotle" are people's *opinions*.

To see this, consider a speaker *S* who speaks language *L*. In order for *S* to know *L*, it must be the case that *S* grasps the senses associated with the symbols of *L*, and correlates them as signs. This will be true of any speaker of *L*, so that if *S* and *S*′ are both speakers of *L* they grasp the same senses, and associate them with the same symbols. If *S* and *S′* are both speakers of *L*, that is, of the same language, and "Aristotle" is a proper name in *L*, then *S* and *S′* associate the same sense with "Aristotle." It should be clear from our discussion thus far that this obtains not only for their knowledge of *L*, but also for their use of *L*; on any occasion of use by *S* and *S′* in any context, "Aristotle" expresses the same sense. Now suppose that we were to inquire of *S* and *S′* what they thought the sense of "Aristotle" is, to give their *opinions* on the matter. *S* says he thinks the sense of "Aristotle" is the pupil of Plato and teacher of Alexander the Great; *S′* says that in his view it is the teacher of Alexander the Great who was born in Stagira. Given that *S* and *S′* have these divergent opinions, perhaps even adamantly so, they will accordingly take it that the sentence "Aristotle was born in Stagira" expresses different thoughts. That they believe that different thoughts are expressed—something that they are wrong about—will often be tolerable in everyday life, so long as they agree that the senses determine the same reference, as this delimits how much people can disagree in their opinions about the sense they attach to the name without breakdown in communication.

Note that on this way of viewing the "On Sense and Reference" footnote, since *S* and *S′* speak the same language, they have opinions about one and the same thing; in that both have the name "Aristotle" in their language, they have each grasped the same sense. They merely *think* what they have grasped is different. How could this be? Consider an analogy. Suppose that two people see an object, although under very disparate perceptual circumstances, and are asked what they have seen. They each provide a distinct description that properly describes the object they have seen, reflecting their different perspectives of the object perceived. It may turn out that these descriptions are so different that neither would think that they are describing the same object. But while each would think that the other had perceived a different object, it would of course be absurd to conclude that there *are* different objects.

With senses, the situation is quite parallel, once we recall that for Frege senses are objects, albeit abstract. So, whereas perception is the cognitive relation in which we stand with respect to concrete objects, the comparable cognitive relation in which we stand to abstract objects is grasping, and what we have just said for concrete objects can be repeated for senses, *qua* abstract objects, with grasping replacing perceiving. Thus, suppose that *S* and *S′* have grasped the sense of "Aristotle," but that they may have done so under disparate contextual circumstances. Upon inquiry as to what the sense of "Aristotle," is, they each describe what they think it is. Their views, their opinions, may turn out to be rather different—the pupil of Plato and teacher of Alexander the Great on the one hand, the teacher of Alexander the Great who was born in Stagira on the other. These are sufficiently different that it might be thought that *S* and *S'* had grasped different senses. But, comparably to perception, it would be a fallacy to conclude from *S* and *S'* thinking they had grasped different senses that there *are* different senses.

As we are reconstructing the footnote from "On Sense and Reference," it is an affirmation of the invariance view of the sense of a proper name. For a given language *L* that contains a proper name, for example, "Aristotle," there is associated one and only sense that it expresses, and any speaker of *L* who has "Aristotle" in their language grasps this sense. Speakers who so grasp this sense have, on this view, opinions that they can express as to what they have grasped; they can say what they think *the* sense is, but nothing requires that they think it is the same. Now in order for speakers to have different opinions, they must have conscious access to the information that is characteristic of the mode of presentation of the sense they have grasped. It cannot be, however, that mere grasp of sense is sufficient for having a sense in mind *in toto*, for then we would not expect speakers to differ in what they think the sense is; if all grasp a given sense in this way, they would all make the same report of what they think the sense is. Access to sense thus may be only partial. But if so, it could then be that speakers differ in what they take the sense to be; the reason that speakers' views as to the sense may differ is that they may access different aspects of the sense. Speakers, as we have it, may bring senses to mind, in the sense of having consciousness of them, although this consciousness may be partial relative to the sense grasped, and divergent from speaker to speaker. The differing descriptions that speakers provide are expressions of this consciousness.

While Frege makes few remarks about the relation we have in consciousness to the senses we grasp, where he does, in his notes of 1914 on "Logic in Mathematics," he seems to eliminate the possibility that we can have senses in mind at all:

> a thought...is in no way to be identified with a content of my consciousness.[30]

If this is Frege's take on the relation of sense and consciousness, it would seem to put the scotch on the reading we have been pursuing. Examination of how Frege reached this conclusion, however, leads to just the opposite conclusion, that Frege is in fact is endorsing the view we are espousing. To see how this is, let us consider Frege's line of reasoning.

---

[30] All quotations from "Logic in Mathematics" (1914), translated by Long and White, in Hermes, Kambartel, and Kaulbach, eds., *Posthumous Writings*, pp. 203–50, on p. 209.

Frege's discussion is placed in the context of making clear that the logical status of thoughts, though related, is to be distinguished from our intuitive notion of thinking; what must be kept clear is the way in which thoughts have "logical significance" as opposed to their "psychological significance." Frege begins explicating this by the following introspection:

> When we examine what actually goes on in our mind when we are doing intellectual work, we find that it is by no means always the case that a thought is present to our consciousness which is clear in all its parts.

"When we use the word 'integral'," Frege continues, "are we always conscious of everything appertaining to it sense?" His answer is negative: "I believe," he says," that this is only very seldom the case." We are limited in our conscious access to sense because our minds are too puny:

> If we tried to call to mind everything appertaining to the sense of this word, we should make no headway. Our minds are simply not comprehensive enough.

Frege then draws the conclusion given above, that "a thought...is in no way to be identified with a content of my consciousness." Given that logical reasoning applies to thoughts, logic cannot be a theory of our conscious thinking. But "nevertheless," Frege tells us, considerations of how we consciously entertain senses is of "great importance for thinking as this actually takes place in human beings."

In reflecting upon Frege's conjectures about thinking in human beings, the key observation is that Frege is not saying that sense *per se* cannot be identified with the content of our consciousness, but only that in general sense *in toto* cannot be. What we cannot reliably bring to mind is a sense in whole. But what is not excluded on Frege's view is that some *part* of the sense may be brought to mind. While we may not be able to get our heads completely around a sense, we might be able to get it around partially. To extend our analogy, just as with perception we will have conscious access to some of the aspects of the object perceived, but not others, with the grasp of sense, we will have conscious access to some of the aspects of the sense, but not others. These aspects will be *proper* parts of senses; they must, as far as a speaker is concerned, uniquely identify a reference, since otherwise the speaker would not believe that they have grasped a sense. But if our conscious access to senses is partial in this way, then what one speaker accesses need not be the same as what another accesses. What one person gets his head around may not be what another person does. Individuals can come to characterize what is grasped in highly disparate ways, depending upon various personal contingencies of grasping (just as with perception). But what this means is that different speakers may come to have differing opinions about the sense they grasped, the very same sense, invariably expressed by a particular sign, and accordingly may believe that different thoughts are expressed. This is just as Frege tells us in the second footnote of "On Sense and Reference." Otherwise, we have no recourse but to conclude that these speakers speak different languages, for their opinions would be of different senses.[31]

V. SENSES ARE NOT DESCRIPTIONS

As I am characterizing Frege's view, it can be summarized as follows: we—those who have grasped a sense—can consciously characterize that sense via a description that is uniquely satisfied by the object presented by that sense, although at best the description provided may characterize the sense only partially. More precisely, the aspect of sense that is given (in part) by the speaker's description is its mode of presentation, and it is about this that speakers may have divergent opinions, as described above.

There is an important implication of Frege's conception that needs emphasis: the descriptions provided by speakers are *not* to be identified with the senses they grasp. This is because *senses are not descriptions.* If they were, an immediate problem would arise, since the descriptions by which senses are given are themselves part of a language. Consequently, they would have to express a sense, which itself, by hypothesis, would also be given by a description, and it too would express a sense, which would be a description. And so on. One might think that this vicious regress could be broken if the sense-giving descriptions had some sort of privileged cognitive status, so that we need ask only what object is (uniquely) presented, that is, what is their reference, and not what sense they express. But then if our analysis of sense-giving descriptions does not allow us to appeal to sense to fix reference, it would seem that the alternative would be to say that such descriptions determine what objects satisfy them by being analyzed in accordance with the theory of descriptions.[32] But reading Russell into Frege in this way would be to give the game away; there is no evidence in Frege's writings to support the view that senses, including those of proper names, are to be given as descriptions that in turn are to be analyzed in accordance with anything like the theory of descriptions.

---

[31] I take it that it is Frege's view that we *can* have consciousness of senses that we have grasped; this is all that is required for his purposes. Yet we may pose a further question that Frege does not consider: is our consciousness of sense either constitutive or exhaustive of our cognitive experience of grasping a sense? If the answer to either is negative, this would imply that there is a further cognitive reality to grasping a sense that is independent of consciousness, some form of tacit knowledge of sense. And although we might hold that this knowledge of sense is inherently partial, we might hold otherwise, that it constitutes knowledge of a sense in its entirety. If so, limitations on our consciousness of sense could be held to reflect how we bring to mind aspects to this knowledge, rather than how we bring to mind directly aspects of the sense itself. But regardless of the inherent interest of this issue, the point remains that grasp of sense, regardless of whether it immediately results in a conscious or tacit state, results in partial consciousness, and this is the important point for the matters at hand.

So far I have considered only the view that the senses of proper names *are* descriptions. A temporizing move might be to hold that proper names do not strictly speaking have senses at all, but gain one by being associated with some definite description (with this association perhaps varying from speaker to speaker, or even utterance to utterance). The sense of the description would then be given in the standard way by the composition of the senses of the constituent concept-words to give a sense whose reference is an object (not a concept). This sort of quasi-Russellian analysis (in the sense that names are "eliminated" in favor of descriptions), would be a way to analyze the senses of names as those of descriptions. The problem is that as a reading of Frege, there is again no textual support that Frege held anything like this view. What he does hold is rather that proper names express a sense *without* any intermediation of descriptions.

If, for Frege, senses are not descriptions, and hence not linguistic nor subject to further analysis, then, one might be curious to know, what are they? Frege himself is quite clear about this: senses are *objects*, albeit abstract objects, that exist independently of any speaker's grasp of them. This is certainly the conclusion we would draw from tracking Frege's linguistic usage, as was observed early on in Frege scholarship.[33] Thus, it was Frege's practice throughout his writings to use a type of locution—"the sense"—whose reference is to an object; as Frege says in "On Concept and Object," "the singular definite article always indicates an object."[34] In "On Sense and Reference," Frege makes this usage explicit. "In order to speak of the sense of an expression 'A'," he remarks, "one may simply use the phrase 'the sense of the expression "A"'."[35] The clear conclusion to draw is that made by Charles Caton,[36] to whom the point here is due: "it follows from the fact that such an expression involves the singular definite article that the sense of any expression is an object."

---

[32] Thus, on Russell's analysis, we do stand in a privileged cognitive relation with the components of descriptions; they are universals with which we are directly acquainted.

[33] An early reference in the English commentary on Frege in which this is observed is Howard Jackson in his "Frege's Ontology," *The Philosophical Review*, LXIX (1960): 394–95.

It is worth noting that Frege allows for the locution "The sense of the expression 'A'" without any of the qualifications that attach to his usage of "the concept *horse*" surveyed in "On Concept and Object." There the issue was that the language requires us to refer to an object, although what we wish to speak of is really something other than an object; hence we may conclude from Frege's unrestricted usage that senses are objects *sui generis*, and do not stand proxy for anything else. In particular, this would indicate that Frege did not take the reference of "The sense of the expression 'A' " to be a proxy object for any sort of function, as value ranges, the references of expressions like "the concept *horse*," are for concepts. (It would be curious if he did. Tempting though it might be to formalize Frege's talk of senses as "determining" or "presenting" as functional mapping onto references, it is hard to see what this mapping would be *from* other than linguistic expressions, so making the existence of senses depend upon that of language.) Thus, given Frege's doctrine that anything that can be referred to is either a function or an object, if senses are not the former, then given that *sinn* can stand as *bedeuting*, not only directly but also indirectly as in propositional attitude contexts, in which signs have references to their customary senses, it follows that senses are objects.[37]

---

[34] On Concept and Object" (1892), translated by Geach, in Geach and Black, eds., *Translations from the Philosophical Writings of Gottlob Frege*, pp. 42–55, on p. 45.

[35] Frege, "On Sense and Reference," p. 59. It is this claim that Russell takes exception to in the Gray's Elegy argument, in which he questions the coherence of the notion of reference being presupposed. Russell argument is that the normal mechanisms of reference, by which it is senses that refer, does not work for referring to senses; rather, what we need for this is an *ab*normal mechanism, in which reference is explicated in some other way than that sense determines reference. But what way this is "remains wholly mysterious" according to Russell, the explication "an inextricable tangle," so much so that the distinction of sense and reference has been "wrongly conceived" and "the point of view in question must be abandoned"—Russell, "On Denoting," *Mind*, XIV (1905): 479–93, at pp. 487–88.

[36] "An Apparent Difficulty in Frege's Ontology," *The Philosophical Review*, LXXI (1962): 462–75, at p. 464.

[37] This argument raises the issue of whether *all* senses are objects, that is, not only senses that have objects as their references, but also senses that have concepts (more generally functions), as references. By the first argument given, since we can speak of "the sense of the concept word 'horse'," it follows that the latter sort of senses are objects too. But note that this implies that the concatenation of thoughts would be

That this was Frege's view is buttressed by a further observation. For Frege, identity criteria are given via abstraction principles, statements of the form $\ulcorner\alpha = \beta \text{ iff } C\urcorner$, where the left-hand side is an identity statement and the right-hand side is a condition over the terms of the identity. Such conditions can only be given for objects; indeed, this follows in the logic of *Grundgesetze*, in which there is only one identity relation and that is objectual identity, so what stands on either side of an identity sign can only be signs for objects. Frege formulates an identity condition for the special case of thoughts, the senses of sentences, in a letter to Husserl from December 1906:

> Now it seems to me that the only possible means of deciding whether proposition *A* expresses the same thought as proposition *B* is.... If *both* the assumption that the content of *A* is false and that of *B* true *and* the assumption that the content of *A* is true and that of *B* false lead to a logical contradiction, and if this can be established without knowing whether the content of *A* or *B* is true or false, and without requiring other than purely logical laws for this purpose, then nothing can belong to the content of *A* as far as it is capable of being judged true or false, which does not also belong to the content of *B* [and vice versa].[38]

---

composed of atomic elements, somehow strung together, and would be unlike the concatenation of sentences, which reflects the relation of the references of their constituents, between an unsaturated concept, and a saturated object. Frege does explicitly remark to the contrary, for instance, as the end of "On Concept and Object," and in the opening paragraph of "Compound Thoughts," where he says that we can "distinguish parts in the thought corresponding to the parts of a sentence, so that the structure of the sentence serves as an image of the structure of the thought"—"Compound Thoughts" (1923), translated by Geach and Stoothof, in Geach, ed., *Logical Investigations*, pp. 55–77, on p. 55. This would appear to imply that concept senses, as well as concepts, are unsaturated, and so cannot be objects. One way of understanding this would be to take concept senses as functions from object senses to thoughts, a view suggested by Geach in his "Critical Notice: *Frege: Philosophy of Language* by Michael Dummett," *Mind*, new series, LXXXV (1976): 436–49, and criticized by Dummett in Chapter 13 of his *The Interpretation of Frege's Philosophy* (Cambridge: Harvard, 1981). The view is also considered in Richard Heck and May in our "Frege's Contribution to Philosophy of Language," in Ernest Lepore and Barry Smith, eds., *Oxford Handbook of Philosophy of Language* (New York: Oxford, 2006); we observe there that this view would not be sufficient to explain why "Cicero is a Roman" and "Tully is a Roman" express different thoughts, unless it were stipulated that the functions in question always map distinct arguments onto distinct thoughts. This leaves us with a quandary. If the senses of predicates are not functions, then given that Frege's ontological distinction is absolute and exhaustive, they must be objects. This is fine, as we would now have a reference for "the sense of the concept word 'horse'," and we would be able to assert the identity of senses of concept words (cf. the discussion of identity criteria immediately below). But it would leave us at something of a loss to say what "unsaturated" means when applied to the senses of predicates.

[38] Letter to Husserl, December 12, 1906, translated by H. Kaal, in G. Gabriel, Hermes, Kambartel, C. Thiel, and A. Veraart, eds., *Philosophical and Mathematical Correspondence* (Chicago: University Press, 1980), p. 70.

While there are matters of interest about the way that Frege formulates this condition,[39] what is more of the moment is that Frege feels that a condition *could* be given at all, for this presupposes that thoughts, and hence the senses of which they are composed, are objects, the only sorts of things for which identity criteria can be given to start with.

Frege's reason, first and foremost, for adhering to the doctrine that senses are objects was his view that this bestowed upon senses their *objective* status. This is brought out in Frege's well-known remarks in the latter part of "The Thought" and elsewhere, including in "On Sense and Reference." In these passages, Frege compares ideas and concrete objects with thoughts, arguing that while thoughts, *qua* sense, share certain properties with ideas and concrete objects, they are not to be identified with either; "a thought," Frege tells us, "belongs neither to my inner world as an idea, nor yet to the external world, the world of things perceptible to the senses."[40] For senses, therefore:

> A third realm must be recognized. Anything belonging to this realm has it in common with ideas that it cannot be perceived by the senses, but has in common with things that it does not need an owner so as to belong to the content of his consciousness.

While this realm is not populated by "the sort of thing to which is usual to apply the term 'actual'," they are nevertheless things with which we have cognitive relations, so that:

---

[39] Note Frege's emphasis that it is the "only possible means." One often hears it said that an identity criterion for senses can be specified such that senses are identical if and only if they can be substituted *salva vertitate* in oblique contexts. Since in such contexts the customary sense is the indirect reference, given Leibniz's Law, if substitutivity in such contexts preserves truth, then the substitutends have the same customary sense. Note, however, that this condition is not a *logical* condition; given the mention of oblique contexts, the conditions cannot be stated within the vocabulary of (pure) logic. In this regard, as an identity criterion on senses it differs from, say, Hume's Principle, Frege's criterion for identity of numbers, which can be stated in this manner. (This is not to say that Hume's Principle is a logical principle, an issue that has been a matter of some considerable debate; cf. George Boolos, "Is Hume's Principle Analytic," in Boolos, *Logic, Logic, and Logic* (Cambridge: Harvard, 1998), pp. 301–14, and Crispin Wright, "Is Hume's Principle Analytic," *Notre Dame Journal of Formal Logic*, XL (1999): 6–30. All I am saying is that given that Hume's Principle can be stated using only the resources of pure logic it is a *candidate* for being a logical principle, something in which a statement that cannot be so given is not.) It is notable that the identity criterion that Frege himself proposes does not explicitly give an identity criterion that presupposes a language with terms for propositional attitudes, and so, unlike that based on substitution in oblique contexts, is a logical condition; if it is satisfied, then, Frege says, "what is capable of being judged true or false in the contents of *A* and *B* is identical, and this alone is of concern to logic, and this is what I call the thought expressed by both *A* and *B*."

[40] Frege, "The Thought," p. 26.

> Outside our inner world we should have to distinguish the external world proper of sensible, perceptible things and the realm of what is nonsensibly perceptible.

To "perceive nonsensibly" an object is just to grasp it; "by grasping a thought," Frege tells us, "I come into a relation with it, and it to me." Entering into such relations with senses is open to all; since thoughts are "independent of me; other men can grasp them as much as I." But this universality just follows from the objectivity of senses, a consequence, Frege thought, of objects, concrete or abstract, being cognitively accessible to all, something that is so regardless of the language anyone speaks.[41]

For Frege, then, the claim that senses are objects amounts to the claim that senses have properties that are characteristic of objects, in particular their independence from speakers and the languages they speak, and their concomitant universal accessibility. Given these properties, in the ontological space Frege assumes—his inclusive ontology of functions and objects—he had little option other than to slot senses into the theory as objects; in this regard, Frege's claim is perhaps best understood as more theoretical than metaphysical. What other answer could Frege give to the question of what senses *are*, given their properties? But to ask this is not the same as asking what senses *do*, or how they are situated cognitively. In answering these questions, descriptions, though to be kept distinct from senses themselves, do have a notable and important role to play. We can take Frege as saying as much in the following remark:

---

[41] Although Frege distinguishes abstract from concrete objects, both are nevertheless "ordinary" objects, in that they can be determined by modes of presentation; thus, numerals express senses that have numbers as their references. But what Frege leaves open about senses is whether they too are ordinary objects, although perhaps it is just that his answer is overdetermined. If he were to abjure from senses, *qua* objects, having modes of presentation, then there could be no appeal to the semantics of sense and reference to explain how we comprehend "the sense of the expression '*A*'." But if this expression has no sense, and so no reference, then whence the support to start with for the inference that senses are objects? Yet abjure we might, especially when we consider senses in a more cognitive light. We may think of modes of presentation as the cognitive states we enter into upon grasp of a sense, that characterize our cognitive perspectives on objects. But while we have perspectives on things, we do not on our perspectives of those things; there is a certain absurdity to having perspectives on our perspectives. Observing this, Akeel Bilgrami and Carol Rovane, in their "Mind, Language and the Limits of Inquiry," in James McGilvery, ed., *The Cambridge Companion to Logic* (New York: Cambridge, 2005), pp. 181–203, draw out as a corollary that modes of presentation are cognitively *transparent*; we know them, as it were, whole, no part can be masked from us, (although it may be from our consciousness; cf. footnote 31). We can be short of knowledge of objects, for if we were not we could not be unaware of identity, that, for instance, that Hesperus is Phosphorus. But how could senses ever be presented to us in such a way that we would not know immediately of their identity, whether the sense of "*A*" is the same as the sense of "*B*"? But if we have no identity puzzles for senses because modes of presentation are transparent to us—"The sense of '*A*' is the same as the sense of '*B*'" would be analytic—then it will also be the case that someone who believes that Cicero is a Roman, but that Tully is not, is in no way uninformed about her beliefs; it cannot be that she has contradictory beliefs but just does not realize it. What she is un- (or mis-) informed about is something about the world, not her beliefs. If we are to conclude from these considerations that senses are not objects, or at least not ordinary objects, it cannot be, however, at the expense of the objectivity and accessibility of senses, or at least of the sort of information that senses are intended to carry; we do not want to give up on the reasons that drew Frege to claim that senses are objects. Arguably, Bilgrami and Rovane maintain, we have the epistemological wherewithall to make good on this, even if Frege did not.

> Although the thought does not belong with the contents of the thinker's consciousness, there must be something in his consciousness that is aimed at the thought...not to be confused with the thought itself.[42]

Here, Frege clearly distinguishes our cognitive awareness of sense from sense itself, and it is natural to think that the descriptions speakers provide express their conscious reflections on the content of the senses they have grasped. So, although the descriptions and the senses are not, as Frege says, to be confused, the former is nonetheless "aimed" at the latter by representing our conscious experience of what has been grasped. Given that a sense is something whose nature is to determine a reference, by giving a description of that reference, of the object presented, we are thereby giving a characterization (albeit partial, yet proper) of a sense. To the extent that we have conscious access to the senses we grasp, how else other than by a description of reference are we to put into words what these senses are? But from this, it would be an obvious fallacy to conclude that senses *are* descriptions; indeed, to be clear, we must be careful not to confuse, according to Frege, senses as objects, and the descriptions by which we, *qua* graspers, characterize their conscious embodiment.[43]

VI. PSYCHOLOGISM VERSUS PSYCHOLOGY

In bringing our discussion to a close, we may ask what ultimately underlies Frege's allegiance to the invariance thesis; why at heart is it of such importance to Frege (as opposed for instance to Russell), that there be only one sense always associated with each expression. Conversely, we may ask what leads Frege to reject the variance thesis in all its forms. The answer is apparent when we consider the context in which Frege discusses the matter. To take an example, in "On Sense and Reference," Frege tells us that "...one need have no scruples in speaking simply of *the* sense."[44] He makes this endorsement of the invariance thesis in the context of contrasting senses, more specifically thoughts, with ideas. Ideas are subjective and variant, essentially so; "in the case of an idea," Frege carries on the remark above, "one must, strictly speaking, add to whom it belongs and at what time." Because of their privacy, ideas cannot be shared, even ideas of the same thing. Senses have just the opposite properties; they are objective and public, they "may be the common property of many and therefore...not a part or a mode of the individual mind." By embracing the invariance thesis, and the concomitant rejection of the variance thesis, Frege is thus resisting the descent of thought into ideas, for it is a mark of ideas, as opposed to thoughts, that they vary from speaker to speaker; variance is a property of ideas, invariance, with respect to their linguistic expression, a property of senses. "The proper means of expression for a thought," Frege tells us, "is a sentence. But a sentence is hardly an appropriate vehicle for conveying an idea." In this regard, "Thoughts, are fundamentally different from ideas (in the psychological sense)."[45]

---

[42] Frege, "The Thought," p. 26.

[43] I am indebted to Bilgrami for illuminating discussion of the issues discussed in this section.

To be sure, Frege continues the comments cited at the beginning of the previous paragraph with a remark that on the face of it might be thought to temporize, if not wholly disown, his allegiance to the invariance thesis. What he says is:

> It might perhaps be said: Just as one man connects this idea, and another that idea, with the same word, so also one man can associate this sense and another that sense (*op. cit.*, p. 60).

But there is no intention with this remark to repudiate the invariance thesis. Rather, Frege means here to allow that *even if* one were to think that thoughts and ideas do not differ in this way, nevertheless it would still be a mistake to identify them. This can be seen from Frege's next remark:

> But there still remains a difference in the mode of connexion. They are not prevented from grasping the same sense; but they cannot have the same idea.

So while it might be conceivable, *counterfactually*, that thoughts and ideas are not differentiated with respect to (in)variance, nevertheless there remains a residual difference between them, in that thoughts are omni-accessible, and ideas are not. Thus, at most with these remarks Frege allows that vis-à-vis ideas, the invariance of thoughts may not be on an equal footing with their accessibility; but this hardly rates as a condemnation of the invariance thesis, or as a sanctioning of its inverse.

---

[44] Frege, "On Sense and Reference," p. 60.

[45] Frege, "Logic," p. 131.

For Frege, the objectivity and publicity of senses is directly tied to their status as objects. But in saying this we nevertheless must be mindful that *qua* objects they are rather queer. This is not because they are abstract.[46] Rather, the peculiarity of senses arises from their being defined relative to other things distinct from themselves, namely, their references (whether they be, given Frege's ontology, functions, including the special case of concepts, of *n*-ary order, or objects, both abstract and concrete, including the True and the False as the references of thoughts, as well as senses themselves, when the references of indirect senses). This characteristic of senses, moreover, extends to the very heart of Frege's conception of logic, since for Frege inferential relations hold not merely between thoughts, but rather only between *true* thoughts, that is, thoughts that have the True as their reference. But peculiar in this regard as senses may be, it is nevertheless essential to them, for that they have this characteristic allows them to play the role that they do as linguistic meanings; grasped as expressed by terms of a language, senses connect symbols to the things signified by those terms.

This aspect of senses, however, raises a persistent worry. If senses are to present objects, then they must do so in virtue of *information* about the objects so presented. But how is this information, which by hypothesis is encapsulated as an objective mode of presentation, ultimately to be kept distinct from our subjective conception of objects? If these cannot be kept distinct, then it would appear that inevitably the former would collapse into the latter, and one would have to admit that senses, in their inherent subjectivity, could vary from speaker to speaker. (So consider the following remark of Graeme Forbes: "I do not deny that the sense of a proper name can vary from person to person, for the sense of the name will express something like a *conception* of its bearer which the subject associates with the name" (*op. cit.*, p. 5).) To admit this would thus place Frege on a slippery slope to a sort of psychologism about senses, while perhaps not as a thesis about the foundations of logic, to wit that the

---

[46] For an illuminating discussion of the role of abstracta in Frege's throught see Teri Merrick, "Frege's Distinction between Concepts and Objects: A Descendant of Kant's Distinction between Concepts and Intuitions" (Doctoral dissertation, University of California, Irvine). Merrick makes the point that Frege's commitment is not to abstract objects *per se*, but rather reflects deeper commitments to scientific theories that incorporate such objects as aspects of their conceptual and empirical structure. Frege's embrace of abstracta is thus relative to the role they play in scientific explanation.

ALIPORE CAMPUS

personal, subjective psychological states of individuals are the objects of reasoning, but as a thesis that such states are the meanings of linguistic expressions. If the descent into psychologism were inevitable, it would minimally create a tension in Frege's thinking, ultimately an unstable one, between the role of thoughts as the objects of reasoning, and their role as objects of judgment.

That such descent is inevitable would seem to follow from the following argument. Suppose that we have (senses with) objective modes of presentation: can they fail to have a reference? If, as we said above, senses are *defined* relative to the objects that are their references, then it would follow that there are no senses without references. From this it would follow that all thoughts would be complete *definite* thoughts—that is, either true thoughts or false thoughts—and this would be so for all languages, logically perfect or not. Continuing the line of reasoning, it follows that the only way that we could have a term in a language that does not have a reference is if it does not express a sense; "Pegasus," on this view, would express no sense, and sentences containing such a term would not express any thought. Since such sentences express no thought, they could neither be involved in logical reasoning nor would they have judgeable content (there would be no grounds for establishing their truth or falsity), and indeed these are, according to Frege, just the properties of sentences containing nonreferential terms But we are left then with a yawning question: Why is "Pegasus flies" even *intelligible*? Not because it expresses a thought; that is, not because of our grasp of objective senses. The alternative is that we can understand such sentences because subjective conditions are satisfied, where we take such conditions as "modes of presentation" that are conceptual in nature, and therefore need not map onto existent objects (we presumably can have a conception of Santa Claus without there being any such thing). But given their subjectivity, such modes of presentation would be private and variable from speaker to speaker, and as such would be indistinguishable, from a cognitive perspective, from ideas. And so, we will have started the slide into the morass of a psychologism of sense, especially so since once the role of subjective modes of presentation is granted, it will become difficult to distinguish cognitively such ersatz conceptual representations from real ones, that is, those that arise from the grasp of sense, and so represent actual references.[47]

---

[47] The argument just surveyed adapts an argument presented by Michael Kremer in his "Sense and Meaning: The Origins and Development of the Distinction" (ms., University of Chicago); cf. Section IX. Kremer develops his argument starting with an

The significance of this argument, were we to accept it, should not be underestimated, for it extends beyond a claim about the psychologism of sense, implying that sense is not to be identified with meaning. Presumably, if a sentence is intelligible to a speaker, so that the speaker can be said to understand that sentence, it is because the speaker knows the meaning of the sentence in question. Given the intelligibility of sentences containing terms without reference, it follows that the speaker knows the meaning of such sentences. But such sentences do not express thoughts; hence grasping the thought expressed by a sentence is not the same as knowing its meaning. Consequently, the characteristically Fregean thesis that it is the grasp of thought that allows a speaker to attain the sort of cognitive state that constitutes understanding a sentence, so that it is the thought that constitutes the meaning of that sentence, is undermined, at least as a general thesis about language.[48]

---

observation initially due to Dummett, in his paper "Truth" (reprinted in *Truth and Other Enigmas* (Cambridge: Harvard, 1978), pp. 1–24). Frege maintains that a truth predicate is eliminable because the statements "'*p*' is true" and "The thought that *p* is true" express nothing more than what is expressed by "*p*" itself. This is not so, however, if "*p*" contains a term whose sense does not determine a reference, since then "*p*", "'*p*' is true" and "The thought that *p* is true" do not express the same thought; "*p*" in this circumstance is neither true nor false, but both "'*p*' is true" and "The thought that *p* is true" are false. Insofar as this is an issue about inference, Frege has a ready dodge, in that logically perfect languages contain no terms without reference. Kremer, however, sees a different problem looming; in languages in which truth is ineliminable as a property of thoughts, Frege's conception of judgment, based on a relation between thoughts and a designated object, will be jeopardized. (There is much to be said about this claim, which we leave for another occasion.) Kremer's response, as per the argument in the text, is that terms without reference are also terms without sense; on this assumption all three of the statements above return to being on par, as being neither true nor false. Kremer then draws out the consequences that modes of presentation, necessary in order to link thoughts to what they are about, and hence to facilitate their judgment, are inherently "private, subjective, and can vary from speaker to speaker," thus committing Frege to a psychologism of senses. Frege's remarks on the variance of sense are to be understood as indicating Frege's acceptance of this characteristic of senses. This creates, according to Kremer, a fundamental underlying tension in Frege's conception of senses with his claims that they are objective and omniaccessible: "The idea of a "common store of thoughts" is threatened when the sense of our words is reduced to the interpretation placed on them by each individual speaker." Kremer is correct in pointing out that this is the consequence of the view he describes; where I think Kremer errs is in holding that Frege held that senses, including their modes of presentation, *ever* have the characteristic subjective privacy of ideas. There is no tension in Frege's views precisely because of his care in distinguishing senses from ideas.

[48] That is, it could still be maintained that in logically perfect languages, in which there are no terms without reference, that thoughts are the meanings of sentences. But now such languages would no longer be continuous with the general conception of

It seems to me that the argument we have been reviewing should be understood in a rather different light than leading to these conclusions. The more compelling way to understand its significance is as a sort of indefinability result; that is, if senses, inclusive of their modes of presentation, are objective, then they are not definable. Definability for Frege, we should note, is not a requirement for objectivity; for some objects, we can at best just elucidate their character by describing how they work and the roles they play. Senses are like this; we can characterize their function in language and thought, but we cannot define them. (Senses, in this regard, are more like chairs than numbers.[49]) If we take this to be the import of the argument, then it does not support the dire conclusions about psychologism and meaning, as there is nothing compelling us to hold that terms without reference are also terms without sense. The elucidation of sense is such that it can be, as Frege puts it, both proper and mock; to terms without reference "we cannot deny...a sense," even though it will not be a "definite sense," nor can we deny sense to sentences that contain such terms, although the thoughts they express will be "mock thoughts," to be contrasted with "thoughts proper, thoughts that are either true or false."[50] Such mock thoughts are sufficient, according to Frege, for understanding a sentence, although it will be understood in a special way, as fictitious. Thus, understood as an indefinability result, the argument can be brought into alignment with what Frege explicitly says about such terms, rather than being at odds with it.

Although we should recognize that nowhere in Frege's writings does it surface that there can be intelligible sentences of a language that do not express a thought, if we were to hold that there can be

---

language, for we would have at least two sorts of languages, those in which thoughts are meanings, and those in which they are not. This would conflict, however, with Frege's view of language; see discussion in section II.

[49] While senses cannot be defined, they can be references (of terms that occur in oblique contexts), and this is sufficient, in Fregean terms, to establish their objectual status. (For discussion of the relation of objectivity and reference, see May, "Frege on Indexicals," and Wright, *Frege's Conception of Numbers as Objects* (Aberdeen: University Press, 1983), chapter 1.) Lack of definability, however, may be a bar to a logic of senses; so, for Frege, it is the definability of numbers, *qua* abstract objects, that brings them within the purview of logic. Note that by indefinability we mean lack of explicit definition, the only sort of definition that Frege countenances. Although we might be able to give identity criteria for senses, (see section V and footnote 39), this would not be sufficient as a definition, given Frege's rejection in *Grundlagen* of identity criteria as contextual definitions in the face of the Caesar problem.

[50] Frege, "Logic," p. 130, except the term "definite sense," which Frege employs in "On Sense and Reference," p. 58.

terms without sense, but that nevertheless have associated (subjective) modes of presentation, we would be committed to holding that modes of presentation are distinct from senses in a way that the question can arise of their objectivity independently of the question of the objectivity of sense itself. (Thus, one who accepted the argument above would distinguish senses as the objects of reasoning and judgment from subjective modes of presentation as the grounds for comprehension and communication.) I do not think, however, that this is correct, as it is based a faulty understanding of the relation of sense and modes of presentation, even though there is an ambiguity in the way their relation is understood. On one way of understanding the notion of mode of presentation, it is employed as part of the elucidation of sense. It describes (at least one way) that sense can be related to reference; it is how we describe, as Frege would have put it in an early terminology, an aspect of the judgeable content of a thought. So understood, modes of presentation are as objective as senses, as they also are on the other way of thinking of modes of presentation, which arises when we consider them cognitively. On this way of seeing things, a mode of presentation is what we get, *qua* cognitive state, when we bring a sense to consciousness. When a sense is brought to consciousness, and thus we gain (partial) access to the content of that sense, we do so in terms of a mode of presentation of the object that is the reference of that sense. Understood this way, a mode of presentation is our conscious experience of a sense; it is the reflection in consciousness of sense. So while with respect to a given expression of a language, all speakers grasp the same sense, nevertheless their modes of presentation, as figments of each individual consciousness, may differ.

On this latter understanding of modes of presentation, it is implied that if there were not a sense grasped, there would be no conscious experience, no mode of presentation; without the grasp of sense, there would not be anything, *qua* language, to have conscious experience *of*. It follows from this that if a sentence is consciously intelligible, that is, understandable, then it *must* express a thought. This is true in the most general sense, in that it applies to all sentences of all languages. In particular, it includes sentences that contain terms without reference, for if we were to deny them senses, we would also have to deny their intelligibility. From the perspective of our conscious experience, then, there will be no difference in the relation of the sense of "Secretariat" to consciousness than the relation of the sense of "Pegasus." The grasp of the sense expressed by one does not give rise to any different sort of state of mind than grasp of the other. But saying that cognitively that there *is* a mode of presentation associated with "Pegasus," is not thereby to claim that its sense contributes to judgeable content; whether this is so is not relevant to whether we can have conscious experience of a sense. It is relevant however, to whether a thought gives grounds for judgment, and so is a definite thought, one that is true or false, and not a "mock" thought. In this regard, nondenoting expressions do not have modes of presentation, in the first sense described above. But this is not something that necessarily reveals itself to our consciousness; we may think that our cognitive modes of presentation are fulfilled, but we may be mistaken in taking fiction to be fact (or vice versa).

Frege was well aware of the danger of the collapse of thoughts into ideas, and it is something against which he inveighed in no uncertain terms, notably in "The Thought" and "On Sense and Reference," as well as elsewhere. His arguments centered around ideas, in their very subjectivity, not having the correct properties to be either the objects of reasoning or the meanings of linguistic expressions, and as such Frege was constructing a bulwark against a descent into psychologism. The invariance thesis is part of this *rejection* of psychologism; but it is a thesis that has appended an associated *psychology*. The psychological issues become most acute when thoughts are considered not so much in their role as objects of reasoning, but rather in their role as facilitators of communication. Thoughts, for Frege, are what make communication possible; *all* languages, from natural languages to the *begriffsschrift*, are useable because they consist of propositions, because their sentences express thoughts that are universally cognitively accessible. In this regard, all languages are of a piece. But given this role—that is, senses grasped as the meanings of linguistic expressions—we must consider the psychology of grasping and communicating thoughts. What we have proposed is that Frege's remarks regarding the putative variances of senses are meant to be understood as constructive parts of this psychology; it is this psychology that takes on the burden of showing how something that is commonly grasped by all can appear to be interpreted differently by different speakers. As such, Frege's remarks are intended in defense of the invariance of sense and thought, and hence of their status (as opposed to our consciousness of them), as the objective meanings of linguistic expressions.

ROBERT MAY

University of California, Irvine

COMMENTS AND CRITICISM

A NEW PROBLEM OF DESCRIPTIVE POWER*

> The woman stared at the fruit....She wanted the wisdom that it would give her, and she ate some of the fruit....So the Lord sent them out of the Garden of Eden.
> —Genesis 3

Analytic philosophers appear to quantify over unactualized possibilities all the time, as if they believed in a plurality of full-blooded worlds, spatiotemporal arrangements of individuals, properties, and relations. For example, metaphysicians discuss whether individuals have some properties in all worlds in which they exist, or whether any pair of properties can coexist in one world. A domain of discourse that includes possibilia allows for sharpening up questions, claims, and arguments. Its benefits are so great that David Lewis called it "Philosopher's Paradise."[1] However, many philosophers are actualists, denying that there are any unactualized entities. Can they nevertheless partake of paradise, that is, make sense of questions phrased in terms of possible worlds and individuals? *Prima facie* it is not clear what actualists are committing themselves to when they affirm an existential quantification over possibilia. They are thus under the obligation to give a story about what they are doing. If they cannot discharge it, their entitlement to use the possibilist language remains in doubt.

I. ACTUALISM AND THE POSSIBILIST PARADISE

Linguistic ersatzism, as characterized by Lewis, seems to offer such a story: the worlds quantified over are not concrete universes, but maximally consistent classes of sentences.[2] Linguistic ersatzism gives truth values of possibilist idioms as a function of the consistency of classes of sentences in a nonpossibilist language. For example, 'There is a possible world in which there is a talking donkey' comes out true because 'There is a talking donkey' is a member of a maximally consistent class of sentences.

---

* Many thanks to Karen Bennett, John P. Burgess, Adam Elga, Alexander Paseau, Daniel Rothschild, Brett Sherman, Ted Sider, and an anonymous referee.

[1] This is the title of part I of *On the Plurality of Worlds* (New York: Blackwell, 1986).

[2] Lewis's influential discussion is in "Paradise on the Cheap?" which is Part III of *On the Plurality of Worlds*. What he calls "linguistic ersatzism" is a technical elaboration of proposals by R.M. Adams, Rudolf Carnap, and Richard Jeffrey.

0022-362X/06/0303/145–62

Does linguistic ersatzism offer a reduction of the possibilist language to a metalanguage, with a consistency predicate, of a non-possibilist language? Lewis has shown that it does not. What is supposed to be the reducing language is too poor in descriptive power to determine the truth values of all sentences of the possibilist language. Lewis's well-known example involves *aliens*, properties not existing in the actual world. He raises the question whether there could be a pair of *role-swappers*: alien properties such that two worlds only differ from each other with respect to which of these two properties plays which role. Linguistic ersatzism, depending on how exactly it is understood, either cannot make sense of this question, or answers it trivially in the negative.[3] This is its so-called "problem of descriptive power." Thus if an actualist aspires to use the full range of the possibilist language, she needs to develop a different theory.

In "The Ersatz Pluriverse," Theodore Sider attempts to do that:

> Like many, I turn to reduction. An adequate reduction of talk of possibilia—worlds and their inhabitants—must be materially adequate, but must not appeal to objectionable entities (like Lewisian worlds). Given such a reduction, the contemporary possible-worlds theorist's tools may be freely used with a clear ontological conscience.[4]

For linguistic ersatzism, the basis of the (partial) reduction is a metalinguistic consistency predicate. It can thus respect the traditional empiricist constraint that all necessity is linguistic, or conceptual. In Quine's terms, linguistic ersatzism is modally involved to the first grade. Sider's theory, in contrast, is modally involved to the third grade. It assumes that not only iterating modal operators, but also quantifying into modal contexts makes sense, and produces sentences which have truth values.[5] The question whether this cost in ideology outweighs the gain in descriptive power is likely to get different answers from different philosophers, depending on how comfortable they are with irreducible *de re* modality and essentialism. I am not trying to adjudicate it here.

---

[3] Unamended, it cannot make sense of the question. It gives a negative answer if 'There is a property X such that there is a world in which $\Phi(X)$' is taken as a substitutional quantification and translated as 'There is a property name N such that '$\Phi(N/X)$' belongs to some maximally consistent class of sentences', where '$\Phi(N/X)$' names the result of substituting N for X in $\Phi(X)$, and where the quantification ranges only over actual names. In that version, linguistic ersatzism implies that there are no alien properties, and *a fortiori* no alien role-swappers.

[4] This JOURNAL, XCIX, 6 (June 2002): 279–315; quotation from p. 281.

[5] Since Sider advocates counterpart theory, he does not assume that for individual variables, but it is crucial for his solution to the problem of alien role-swappers that quantification over property variables into modal contexts is allowed.

Sider is explicit that his theory is reductive specifically about the possibilist language, and does not attempt a reduction of the modal notion of possibility as well. If the reduction is successful, any dispute in the possibilist language is meaningful in the reducing language. Sider claims that the theory he presents, which I call *pluriversal ersatzism*, achieves such a reduction. He lists four examples of possibilist sentences that actualists might want to use, and argues:

> The reduction does indeed assign truth conditions to these sentences.... Thus my reduction allows one to partake fully of modal metaphysics and semantics (306).

However, the last statement is a *non sequitur*; the reduction might succeed with these four examples but fail with others. Thus it has not been shown that Sider's theory gives actualists a license to use the full possibilist language. I will show that in fact, it does not give such a license, by presenting an example of a possibilist distinction Sider's theory cannot cope with. The theory thus fails to do the job it is designed for.

II. PLURIVERSAL ERSATZISM: EXPOSITION

Pluriversal ersatzism uses a modal language, which I call *ML*, as its reducing language. A reduction does not need to provide a recursive translation between two languages. It can make a detour through model-theoretic semantics, if the same models are used to define truth of sentences of the two languages. In the case at hand, so-called "modal models" are the relevant ones, set-theoretic structures modelling the space of possible worlds (292–93). Possibilist sentences can straightforwardly be interpreted as true or false in one of these models. Sentences of *ML* are evaluated in the way familiar from the possible-worlds semantics for modal logic. If every *ML*-sentence that is true (*simpliciter*) is true in a modal model $\mathcal{M}$, then $\mathcal{M}$ is *realistic.* Sider claims:

(1) A possibilist sentence is *true* iff it is true in every realistic model; it is *false* iff its negation is true.

A simple example may illustrate how the reduction is supposed to work. Suppose that 'It is possible that there is a talking donkey' is a true *ML*-sentence. This sentence is true in all and only the modal models in which for some *w*, the extension of 'talking' in *w* and the extension of 'donkey' in *w* overlap. Since the sentence is true, every realistic model fulfills that condition. Hence the possibilist sentence 'There is a world in which there is a talking donkey' is true in every realistic model. By (1), it is true. If we suppose, on the other hand, that 'It is possible that there is a talking donkey' is false, we can conclude by similar reasoning that 'There is a world in which there is a talking donkey' is false. In either case, the theory succeeds in giving a truth value to that particular possibilist sentence.

Does it likewise succeed for all other possibilist sentences? This crucial question about the expressive power of *ML* is not addressed in full generality in Sider's article. As I noted above, he claims in one place that the reduction works, based on the successful handling of a few examples. However, in another place he acknowledges that he has not proved this: "It is unclear to me at present whether a harmful multiplicity arises from CONSTRAINTS failing to constrain realistic models up to isomorphism."[6] Before showing that such a multiplicity may arise, I want to clarify what exactly a reduction would have to achieve.

It is crucial that for any sentence in the possibilist language, it can be shown that it has a truth value, independently of what the *ML*-truths are. In the above example, this condition is fulfilled; we do not have to assume either the truth or the falsity of 'It is possible that there are talking donkeys'. It is a language that is to be reduced, not a particular theory formulated in that language. Sider rightly points out in several places that the reduction should not be hostage to substantive claims.[7] It is supposed to be neutral among competing metaphysical views, since it only provides a framework in which they can be debated. For all we know, linguistic ersatzism might give the right verdict on alien role-swappers, since for all we know there might be no alien properties at all.[8] But for all we know, it gets it wrong, and that is enough to disqualify it as a reduction.

In logician's terminology, two models are *elementarily equivalent* relative to a language *L* if exactly the same sentences of *L* are true in both models. Given that isomorphism between two models is defined in such a way that no possibilist sentence has different truth values in isomorphic models, reducibility would be established if it could be shown that only isomorphic modal models are elementarily equivalent relative to *ML*. On the other hand, reduciblity can be shown to fail by describing, in the possibilist language, a pair of nonisomorphic modal models that are elementarily equivalent. This is, in effect, what Lewis did for linguistic ersatzism, and what I will do for pluriversal ersatzism.[9]

---

[6] Footnote 27, p. 295. 'CONSTRAINTS' denotes the class of true *ML*-sentences that are *de dicto* with respect to quantifiers binding individual variables.

[7] For example, on p. 286 and pp. 314–15.

[8] Linguistic ersatzism can be taken to say that there are no alien properties. See note 3.

[9] While section III gives an informal argument, the appendix presents a proof.

Whether the possibilist language is reducible in the sense explained depends on the expressive power of *ML*. Sider makes this language, which has a possibility and an actuality operator as modal primitives, very powerful. It has a name for every actual individual and property. Since every actual entity is its own name, it is a so-called *Lagadonian* language.[10] Moreover, *ML* allows infinite disjunctions and conjunctions. The quantifiers are world-restricted, or actualist: when the modal-model-semantics evaluates a formula at w, it interprets the quantifiers as ranging only over the domain of w (292–93).

ML is supposed to be the reducing language. I have so far only vaguely characterized the language to be reduced as the "possibilist language" that philosophers are using. It is ostensibly talking about individuals instantiating properties at worlds.[11] Fortunately, we do not need to specify a precise syntax for it. To ask whether it is reducible, it suffices in any case to know its expressive power in terms of modal models. The question then becomes whether the modal language can express the same distinctions among these models.

When arguing that there are nonisomorphic modal models that are elementarily equivalent, I will use our possibilist language, which is a fragment of what is sometimes called "philosopher's English," that is ordinary English plus logic and set theory. Thereby I will show that the two models are distinguishable in the possibilist language, and hence that the reduction fails.

We can thus ignore in the following the second Lagadonian language that Sider's paper introduces. It is an infinitary possibilist language, which is also given a semantics in terms of modal models. For the purpose of reduction, its details are irrelevant. The reducing language ML may certainly be a language we cannot speak. But that some language can be reduced is relevant only for those who want to speak that language. Hence a reduction of Sider's Lagadonian possibilist language would be relevant only for the philosophers of Lagado, of whom Captain Lemuel Gulliver reports the following:

> [M]any of the most learned and wise adhere to the new scheme of expressing themselves by *things*, which hath only this inconvenience attending it, that if a man's business be very great, and of various kinds,

---

[10] A qualification is needed here: if every actual entity were its own name, then some symbol would have to be ambiguous, for we need logical expressions besides names in the language. The problem can be solved by a code for assigning sets as non-Lagadonian names for other sets, thereby making some small sets available to serve as logical expressions.

[11] Only ostensibly for an actualist, of course. A realist about possible worlds, like Lewis, interprets it literally.

> he must be obliged in proportion to carry a greater bundle of *things* upon his back, unless he can afford one or two strong servants to attend him. I have often beheld two of these sages almost sinking under the weight of their pack.[12]

But along with "the women in conjunction with the vulgar and illiterate" of Lagado (*ibid.*, pp. 198–99), non-Lagadonian philosophers will refuse to adopt the new scheme, and the question whether Sider's Lagadonian possibilist language is reducible is of no direct interest to them.[13] While the language to be reduced should be one that we speak, but may not be able to understand independently of any translation, the reducing should be one that we understand independently, although it may not be one that we speak. This last point will be important later, in the assessment of proposals to get around the problem of *ML*'s lack of expressive power.

III. PARADISE REGAINED?

Sider's theory fails to reduce the possibilist language to a nonpossibilist modal language. A reduction needs to answer all questions that can be asked in the possibilist language, not just those which Lewis used as examples in his critique of linguistic ersatzism. The case of role-swapping has a feature that makes it easily amenable to be settled by *ML*: the pair of worlds whose existence is in question is stipulated to have the same domain of properties. Since in *ML*, the quantifiers in each world range only over entities existing at that world, we might suspect that trouble arises when we ask whether there are worlds with structurally identical descriptions but different domains of properties.

Like the old problem of descriptive power, the new problem arises from distinct entities playing the same role. I say that properties X and Y are *alternatives* of each other iff for every world w there is a world v that has exactly the same pattern of property instantiation, except that w has Y in all places where v has X.[14] If we wish, we may think of properties having an essence which determines what roles they possibly play. On that picture, properties are alternatives if they have the same essence. Among alternatives, there are both role-swappers and *substitutes.* Role-swappers coexist in a world and stand in a certain relation to each other in that world. Since they are alternatives, there is a different world where the converse relation holds between them. But if alternatives are not compossible, not coexisting in any world, they are substitutes.[15] Two teammates who can both play either central midfield or right midfield are role-swappers, and two others who can only play as goalkeepers are substitutes of each other, unable to play together in the team.[16] Sider's theory can deal with role-swappers, even if they come in infinite numbers, thanks to its use of an infinitary language. It can deal with finitely many substitutes as well. But as I now show, it cannot reduce sentences about an infinite cardinality of substitutes, which are very easily formulated in a possibilist language, for example by the sentence 'There are uncountably many substitutes'.[17]

---

[12] Jonathan Swift, *Gulliver's Travels* (New York: Barnes and Noble, 2003), p. 199; originally published 1726.

[13] It can be relevant only indirectly, via some relation between its expressive power to the expressive power of a language we do use. By Sider's Theorem on p. 296, the Lagadonian possibilist language is reducible if no two nonisomorphic modal models are elementarily equivalent relative to *ML*. This leaves the question how our possibilist language compares in expressive power to the Lagadonian possibilist language. Surely, there are things expressible in that language that are inexpressible in our finitary language, where many actual things do not have names. On the other hand, it seems safe to assume that everything expressible in our language can be expressed in the Lagadonian possibilist language. However, my argument will not depend on settling this question.

[14] If X does not exist in w, then w and v are identical.

How can it be expressed in ML that there are at least two alien substitutes? Let the open sentence $\Phi(X)$ be a *repertoire* for a property, specifying all the complete roles a property can play, and saying that they are all the roles it can play. If C is a class of complete world descriptions W in one free variable, $\Phi(X)$ is of the following form ($\bigwedge$ and $\bigvee$ take classes of formulas as arguments and form the conjunction or disjunction, respectively, of all their members):

$$(2) \qquad \bigwedge\{\Diamond W(X)\colon W \in C\} \wedge \Box[\exists Y(Y = X) \rightarrow \bigvee\{W(X)\colon W \in C\}]$$

For any property, it is noncontingent whether it satisfies $\Phi(X)$.[18] Suppose the repertoire is such that two properties having it cannot coexist. Thus the following is a true sentence of *ML*, implying that whatever Φ-alternatives there are will be substitutes, not role-swappers:

---

[15] Alternatives can be role-swappers with respect to some roles and substitutes with respect to other roles, but this does not matter to my argument.

[16] We can think of two worlds with substitutes as differing from each other merely quiddistically, by a "brute" nonidentity of properties. Alternatively, we might think of them as qualitatively different from each other, but in ways that are ineffable from the vantage point of this world.

[17] Like Lewis, Sider takes individuals to be world-bound, and he builds this assumption into the definition of a modal model. World-bound individuals cannot be role-swappers. However, they may be substitutes, and the ensuing argument might as well be put in terms of individuals as of properties. As an anti-haecceitist, Lewis classifies worlds that differ only through substitution of individuals as indiscernible, but as a quidditist, he classifies worlds that differ only through substitution of properties as discernible. (I presume that Sider agrees with Lewis on this issue.) I put the argument in terms of properties rather than individuals to get the stronger conclusion that *ML* fails to distinguish between different cardinalities of discernible worlds.

[18] Following Sider, I am assuming S5 here. Otherwise a conjunction of sentences that start with a modal operator, such as (2), may be true in one world and false in another.

$$\Box \forall X \forall Y (\Phi(X) \wedge \Phi(Y) \rightarrow Y = X) \tag{3}$$

This assumption about Φ is dispensable for my argument, but it makes its formulation easier. For example, Φ might say that whenever X is instantiated, it is the only property of a *loner*, an individual not coexisting with anything contingent except its own parts. (4) expresses that some possible property with repertoire Φ has at least one alternative:

$$\Diamond \exists X \Diamond \exists Y (X \neq Y \wedge \Phi(X) \wedge \Phi(Y)) \tag{4}$$

Following that pattern, we can express that some possible property with repertoire Φ has at least two alternatives:

$$\Diamond \exists X \Diamond \exists Y \Diamond \exists Z\, (X \neq Y \wedge X \neq Z \wedge Y \neq Z \wedge \Phi(X) \wedge \Phi(Y) \wedge \Phi(Z)) \tag{5}$$

It is now obvious how to express, for a given natural number n, that some property with repertoire Φ has at least n alternatives. By conjoining (3), we can express that there are at least n substitutes. Let the sentence expressing that be Φ_n. A class of sentences that contains Φ_n, for every n, only has models with infinitely many Φ-substitutes. Thus *ML* can express that there are infinitely many substitutes. So far, all is well with *ML*.

But this language cannot distinguish among infinite cardinalities of substitutes. To be sure, *ML* allows infinite blocks of quantifiers and infinite conjunctions. But this does not yield infinite alternations of quantifiers with other symbols. In standard infinitary languages, there are no infinite alternations of existential and universal quantifiers, or of quantifiers and negations. Likewise, there are no infinite alternations of diamonds and quantifiers in *ML*; the pattern at the beginning of (5) can only be repeated finitely many times. Thus a sentence built after this template can distinguish only between models with finite numbers of alternatives. The position of the Φs and the negated identities may be different than in (5), but any sentence expressing that there are n alternatives must alternate quantifiers and modal operators no fewer than n times. Since sentences of that type are the only candidates, no *ML*-sentence can make any distinctions among infinite cardinalities of them.

Suppose now there are infinitely many substitutes for some possible property. Then for any infinite cardinal κ, there will be a realistic modal model in which there are exactly κ substitutes. Hence all sentences in the possibilist language about infinite cardinalities of substitutes will lack a truth value, given (1). As a consequence, *ML* cannot tell us how many individuals, properties, or worlds there are in modal space. There will be realistic modal models of different cardinalities, and some claims about the cardinality of the class of worlds or of the class of properties will lack a truth value.

I anticipate the response that the new problem of expressive power is not serious. In practice, it may be argued, philosophers do not make claims whose truth depends on how many substitutes there are.[19] In my view, it would be good if this were true, but I am not sure it is. For once one is thinking in the possible-worlds framework, such claims appear natural. This can be illustrated by an example involving worlds that are structurally like ours, that is have the same pattern of distribution of properties in regions of space-time, but possibly with different properties in corresponding places. In some world that is structurally alike, schmass may be instantiated wherever mass is in the actual world.

Suppose a hypothetical philosopher Louise thinks there are worlds structurally like ours where mass, but no other property is replaced. Since any finite upper bound on the number of such replacement worlds would be arbitrary, Louise claims that there are infinitely many of them. She holds that all these replacements have the same repertoire. Slightly heretically, she denies that mass and its replacement properties are freely recombinable in such a way as to be able to coexist in a world. Hence she is committed to there being infinitely many substitutes of mass. Further, she holds that we cannot epistemically rule out any world that is structurally like our world, and that our credence in a world that we cannot rule out is nonzero. From these claims it follows that at least as many worlds are given nonzero credence as there are substitutes of mass. Since at most denumerably many worlds can be given nonzero credence,[20] Louise is committed to affirm the two possibilist sentences 'There are infinitely many substitutes of mass' and 'There are not uncountably many substitutes of mass'. But according to pluriversal ersatzism, if the first of these is true, the second lacks a truth value, and if the second is true, then the first is false.[21] Hence if pluriversal ersatzism is right, these sentences are not both true, and thus neither is Louise's theory, taken as a set of sentences formulated in the possibilist language. By issuing this verdict, pluriversal ersatzism clearly violates its professed neutrality among philosophical theories that have a consistent formulation in the possibilist language.

---

[19] This response lowers the standards for a successful reduction. It is then not at all clear any more whether linguistic ersatzism does not meet the lower standards as well, for philosophers do not too frequently make claims whose truth value depends on the possibility of alien role-swappers either.

[20] A world has positive probability iff for some n, $P(w) > 1/2^n$. Let $W_n = \{w : P(w) > 1/2^n\}$; then the set of all worlds with positive probability is $\bigcup_{n \in N} W_n$. But for a given n, W_n has at most 2^n elements, since the probabilities cannot sum to more than one. Hence the set of worlds with positive probability is a countable union of finite sets, and thus itself countable.

IV. CONSIDERING PATCHES

My argument shows that *ML*, the language Sider specified, is not powerful enough to reduce the possibilist language. The question remains whether something in the spirit of his proposal does work. Is there some technical trick that could save the project? I offer reasons to think that this is unlikely. The limitations of *ML* may well be the limitations of any approach that takes the modal operators as primitive.

To begin with, it is important to be clear what the rules of the game are. In one sense, it is trivial to express that there are uncountably many substitutes in a language of boxes and diamonds. For we can express it in English, and we could easily set up a coding of sentences of English in a language of boxes and diamonds (or, for that matter, any other language with a sufficiently rich syntax). We could then stipulate that the sentences of that language mean what their English decodings mean. Of course, though, this is of no significance to the question whether the possibilist language can be reduced. It is a constraint on the reduction that the box and diamonds have, at least roughly, the meaning of our familiar 'necessarily' and 'possibly'.

Still, we need to ask whether the power of the language can be increased while being faithful to what the symbols mean. In this section, I examine two pertinent proposals. One is technically attractive, but does not afford a reduction in a philosophically interesting sense, as I argue in the next section. The other can be dismissed because it surreptitiously introduces a possibilist quantifier.

The first proposal involves a radical departure from the usual way of specifying the syntax of infinitary languages: a modal language with non-well-founded, or infinitely deep, sentences. To introduce this notion, it is useful to think of a sentence as represented by a tree. For example, $\exists x\ (Fx \wedge Gx)$ can be represented by a tree of length 3; its bottom node is labelled by $\exists x$; above it, there is a node labelled by $\wedge$; on level 3, there are two branches, their nodes labelled by Fx and Gx, respectively. Standard languages only allow trees that are both of finite width and length. Infinitary languages of the type Sider employs make room for infinitely wide trees. Their formation rules allow $\wedge$ and $\vee$ to apply to infinitely many formulas, and the quantifiers to a formula and infinitely many variables. Infinitely deep, or non-well-founded sentences can be represented by trees that are infinitely long.[22] Since my argument above relies crucially on *ML* not being able to alternate $\Diamond$ and $\exists$ infinitely many times, infinitely deep languages promise to remedy the problem of expressive power. In the attempt to give a semantics for infinitely deep languages, concepts and methods from game theory proved to be fruitful.[23]

---

[21] Suppose 'There are infinitely many substitutes of mass' is true. By the argument above, there is a realistic model with countably and a realistic modal with uncountably many substitutes of mass. By (1), 'There are not uncountably many substitutes of mass' is neither true nor false. Suppose that latter sentence is true. Then there are at most countably many substitutes of mass in every realistic modal model. By the argument above, there are at most finitely many; hence 'There are infinitely many substitutes of mass' is false.

I am now sketching a game-theoretic semantics of an extension *IDML* of *ML* that allows infinitely deep formulas.[24] A sentence and a modal model together define a game with two players, Myself (the initial verifier) and Nature (the initial falsifier). If the game reaches a true atomic formula, I win. If the game reaches a false atomic formula, Nature wins. A sentence is true in a model if I have a winning strategy, and false if Nature has a winning strategy.[25] A game starts at the bottom node and then moves up the tree. *IDML* has the standard rules of game-theoretic semantics: at a conjunction node, the falsifier chooses a branch; at a disjunction, the verifier chooses a branch; at a negation, the roles of the verifier and the falsifier are reversed; at an existential quantification node, the verifier chooses an element of the domain of the world of evaluation. At $\Diamond$, the verifier chooses a world in the modal model.

---

[22] Clearly, such sentences cannot be formed in *ML*, as specified by Sider: "The language is infinitary, so the usual truth conditions for infinite conjunctions and infinite blocks of quantifiers must be adapted to the modal case in the natural way" (293). Sider refers us to M.A. Dickmann, *Large Infinitary Languages: Model Theory* (Amsterdam: North-Holland, 1975), a work that does not discuss non-well-founded languages.

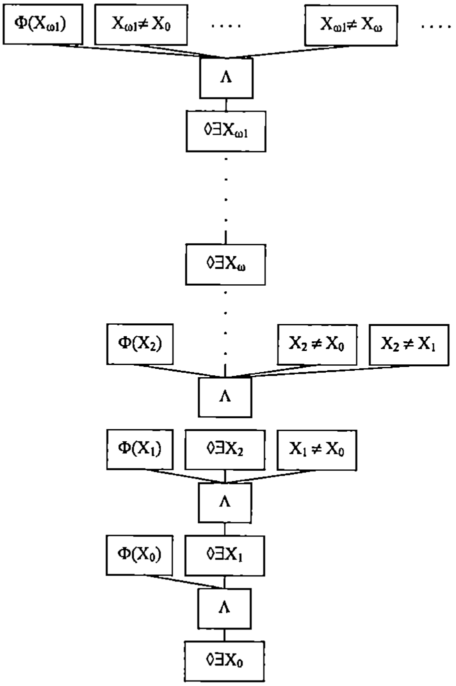
[23] The original and most widely used example of an infinitely deep construction is provided by the *Henkin quantifier* or *game quantifier*, which is equivalent to an infinite alternation of universal and existential quantifiers. It is given a game-theoretic interpretation in Yiannis N. Moschovakis, "The Game Quantifier," *Proceedings of the American Mathematical Society*, XXXI, 1 (January 1972): 245–50. Infinitely deep languages also allowing infinite alternations of $\vee$ and $\wedge$ have been explored by Jaakko Hintikka and Veikko Rantala, "A New Approach to Infinitary Languages," *Annals of Mathematical Logic*, X (1976): 95–115.

[24] For an introduction to game-theoretic semantics, see, for example, Hintikka and Gabriel Sandu, "Game-Theoretical Semantics," in Johan van Benthem and Alice ter Meulen, eds., *Handbook of Logic and Language* (Cambridge: MIT, 1997), pp. 361–410.

[25] The notion of a strategy may be understood in an intuitive way; formally, it is a function mapping any initial segment of the game onto a next move.

The infinitely deep sentence *S*, represented by the figure, is only true in modal models with uncountably many alternatives with repertoire Φ. It is a tree of length $\Omega_1 + 4$, where Ω_1 is the first uncountable ordinal.[26]

Suppose the game *S* is played over a modal model $\mathcal{M}$ with uncountably many Φ-substitutes. Let it be my strategy that at each diamond, I choose a different world, and at the existential quantifier I choose a property with repertoire Φ. If Nature at some point chooses a finite branch, leading to an atomic formula, she loses. Since sooner or later she has to do so, I always win on this strategy, and consequently *S* is true in $\mathcal{M}$.



---

[26] To save space, I draw one node with $X_1 \neq X_2$ instead of one node with a negation sign below a node with $X_1 = X_2$, and I do not draw separate nodes for $\exists$ and $\Diamond$.

I choose a sentence of the length of a successor ordinal, to avoid discussion of what constitutes winning a game that ends at a limit ordinal stage; see Hintikka and Rantala.

Suppose now that S is played over a modal model $\mathcal{M}'$ with at most countably many Φ-substitutes. Then it cannot be the case that I choose a world and a property that both has repertoire Φ and is different from any previously chosen property uncountably many times. As soon as I fail to do so, Nature may choose the appropriate finite branch. Hence Nature has a winning strategy, and S is false in $\mathcal{M}'$.

A suitable elaboration of this sketch of *IDML* may well technically yield a reduction of the possibilist language. However, given the story of punctures and patches so far, it seems reasonable to remain skeptical until we have a proof that such a language can characterize modal models up to isomorphism. However, I argue in the next section that even if such a proof were forthcoming, its philosophical significance would be limited. Before that, I want to show that the second proposal, requiring a less drastic revision of syntax and semantics, does not help.

In my argument above, I exploit the inability of *ML* to alternate diamonds and existential quantifiers infinitely many times. Thus there is a temptation to modify the language to allow an infinite repetition of the pattern $\Diamond\exists$, maybe using the notation $[\Diamond\exists]$.[27] Thus if Φ is a formula and $\tau = v_1, v_2, \ldots$, a sequence of variables, $[\Diamond\exists]\tau\Phi$ should be read as $\Diamond\exists v_1, \Diamond\exists v_2, \ldots, \Phi$.

Since *ML* already allows infinite repetitions of $\exists$, this might seem a small step. However, there is a crucial difference between the ways $[\Diamond\exists]$ and $\exists$ operate on infinitely many arguments. $\exists$, in effect, takes in a set, and the result of applying it can be evaluated in one step. $[\Diamond\exists]$, on the other hand, takes in a sequence, and the result of applying it needs to be evaluated in as many steps as there are elements in the well-ordering. For every occurrence of $\Diamond$ takes us to another world in the model. As long as there are only finitely many variables, this presents no problem, since then the result of pre-fixing $[\Diamond\exists]$ is equivalent to a formula in the language without that new symbol. But in order to increase the expressive power, $[\Diamond\exists]$ needs to operate on an infinite sequence. Semantically, the resulting sentence is infinitely deep, or non-well-founded; there is an infinitely descending chain under the relation of being a proper subformula. In the syntax, this is merely disguised by the use of the notation $[\Diamond\exists]$. We are used to evaluating a formula in finitely many steps until we reach atomic formulas. Infinitely deep formulas do not allow this. To interpret them, we may still proceed in steps indexed by the ordinals; but the familiar rules leave open how to proceed at limit ordinal steps.

---

[27] Thanks to Sider and an anomymous referee for suggestions on this point.

If this proposal is to be also semantically and not just syntactically different from the game-theoretic one, another rule for the limit ordinal step is needed. It is tempting to simply stipulate that after $[\Diamond\exists]\sigma$, where σ is a sequence of variables whose length is a limit ordinal, Φ is to be evaluated at the actual world. Thus $[\Diamond\exists]\sigma\Phi$ is true in a modal model if there is a world w_1 where a σ_1 exists and a world w_2 (that may or may not be distinct from w_1) where a σ_2 exists and... such that Φ is true of σ_1, σ_2, and so on, at the actual world. Now the language does indeed have the power to express that there are uncountably many substitutes of a given property. But this has no tendency to help Sider's project of a reduction that is acceptable for actualists. For $[\Diamond\exists]$ with these stipulated truth conditions is just a cumbersome notational variant of an infinitary possibilist quantifier, that is, a quantifier that ranges at the actual world over all possibilia, whether they exist in that world or not. It is a quantifier of the sort that an actualist is seeking a reduction for.

I do not know how to show in general that no other such move will help, but I am extremely skeptical. We have to recognize that the diamond and the existential quantifier already have their meaning given by their recursive clauses. It is not open to us to stipulate that they have special meanings when they happen to occur in a certain syntactic configuration. In the above example, the notation $[\Diamond\exists]$ tried to suggest that we have a combination of familiar symbols, but failed to conceal that it was a new primitive operator in the language. I cannot pre-empt every actualist attempt to introduce a new operator and argue that we understand it independently of the possibilist language. But before such a proposal is made and adequately defended, possibilist discourse cannot count as reducible in a way that is acceptable to actualists.

V. RENOUNCING PARTS OF PARADISE?

My argument shows that pluriversal ersatzism fails by Sider's standards. A reduction is not supposed to rule on substantive questions in modal metaphysics. In particular, it should not rule that there are most finitely many alternatives, in order to prevent embarrassing questions from arising. Thus it needs to make sense of all questions asked in the possibilist language.

By adopting *IDML*, that problem might be overcome. This language has a deviant syntax, though, and calling it a "language" at all might be stretching the notion. But even if a solution to the technical problem can be found, there remains a philosophical worry. A reduction is only philosophically (as opposed to technically) significant if the reducing language is understood independently of the language to be reduced. There is a serious worry that already *ML* cannot be understood independently of the possibilist language. The problem is not that we do not speak a Lagadonian language, but rather that it is not clear that we have a conception of what it is for sentences about role-swappers and substitutes to be true or false, if we are not thinking in terms of possibilia. For the sake of the argument, however, I grant that we do. But I want to insist that *IDML* would be a philosophically ill-motivated choice of a reducing language. A language with sentences that cannot be produced by any finite number of applications of rules is essentially beyond what finite creatures like us can comprehend. If we cannot comprehend it, there is no reason to assume that it is indeed a language, in the sense of a class whose members are true or false, and not just some nonrepresenting mathematical structure. As far as I can see, the only way in which we could specify an interpretation for *IDML* is by using the possibilist language, which would defeat the purpose of the reduction.

I recommend that actualists brace themselves for the discovery that possibilist discourse is not fully reducible to modal sentential operators. This need not tell against actualism, but rather against some instances of possibilist discourse. Maybe our hypothetical Louise should stop worrying about how many substitutes mass has. The path of wisdom is to start with the modal distinctions that do make sense, and tailor the use of the possibilist language accordingly. By going that path, we might well renounce parts of paradise, but *eo ipso* rid ourselves of *Scheinprobleme* raised by some unanswerable questions that can be asked in the possibilist language. As for renouncing the comforts of paradise, philosophers should be ready to pay that price for attaining wisdom.

STEPHAN LEUENBERGER

Princeton University

APPENDIX: PROOF OF ELEMENTARY EQUIVALENCE

This appendix proves the elementary equivalence of two non-isomorphic modal models relative to the modal language. $\mathcal{M}$ has nondenumerably many worlds. In each of them, exactly one individual, one monadic property, and the two-place identity relations for individuals and for properties exist (there are no higher-order properties and relations), and the individual instantiates the property. The individuals and properties of all these worlds are distinct. The worlds of $\mathcal{M}'$ are a denumerable subset of those of $\mathcal{M}$; the actual world @ is the same in both models. (It would be easy to represent $\mathcal{M}$ and $\mathcal{M}'$ as 6-tuples corresponding to Sider's definitions on pp. 292–93.)

The language, here called L, has names a for the actual individual, P for the actual property, and $=$ for the identity relation.

Let a mapping $*$ of formulas of L into formulas be defined recursively as follows:[28] (i) For atomic ϕ, $\phi = \phi^*$. (ii) $(\neg\phi)^* = \neg(\phi^*)$. (iii) $(A\phi)^* = A(\phi^*)$. (iv) $(\bigwedge\Phi)^* = \bigwedge(\Phi^*)$. (v) $(\Diamond\phi)^* = \Diamond(\phi^*)$. (vi) $(\exists \vec{x}\, \phi(\vec{x}))^* = \exists x_0\, \phi^*(\vec{x_0})$. (vii) $(\exists \vec{X}\, \phi(\vec{X}))^* = \exists X_0\, \phi^*(\vec{X_0})$. ($\Phi^* = \{\phi^*: \phi \in \Phi\}$; $\vec{x}$ is a (possibly infinite) sequence of individual variables; $\vec{x_0}$ is a sequence of the same length, each of whose entries is x_0; and analogously for the upper-case property variables.)

> *Lemma 1*. (a) For every L-formula ϕ, world w and assignment σ, $\mathcal{M}_{w\sigma} \models \phi$ iff $\mathcal{M}_{w\sigma} \models \phi^*$. (b) The same for $\mathcal{M}'$.

Proof. This is trivial for atomic formulas. The inductive steps for negation, conjunction, the actuality-operator and the diamond are straightforward. Existential quantification: suppose $\mathcal{M}_{w\sigma} \models \exists\vec{x}\phi(\vec{x})$. Since $\exists\vec{x}\phi(\vec{x})$ is logically equivalent to $\exists\vec{x}(\bigvee \{x_0 \neq x_i\text{: } i \neq 0 \text{ and } x_i \text{ occurs in } \vec{x}\} \wedge \phi(\vec{x})) \vee \exists x_0\phi(\vec{x_0})$, $\mathcal{M}_{w\sigma} \models \exists\vec{x}(\bigvee \{x_0 \neq x_i\text{: } i \neq 0 \text{ and } x_i \text{ occurs in } \vec{x}\} \wedge \phi(\vec{x}))$ or $\mathcal{M}_{w\sigma} \models \exists x_0\phi(\vec{x_0})$. Since the first disjunct is false in $\mathcal{M}$, $\mathcal{M}_{w\sigma} \models \exists x_0\phi(\vec{x_0})$. Let $\sigma(x/w)$ denote the assignment that differs from σ at most in that it assigns the individual in w to x. Hence $\mathcal{M}_{w\sigma(x_0/w)} \models \phi(\vec{x_0})$, and by the induction hypothesis, $\mathcal{M}_{w\sigma(x0,w)} \models \phi^*(\vec{x_0})$. Hence $\mathcal{M}_{w\sigma} \models \exists x_0\phi^*(\vec{x_0})$. The argument for the property quantifier is exactly analogous; and so is the induction to prove (b).

The language L^* is defined like Sider's L, except for the quantifier- and the conjunction-rule: if ϕ is a formula of L^*, then $\exists x\, \phi$ and $\exists X\, \phi$ are formulas of L^* (however, infinite blocks of quantifiers cannot be introduced); and if Φ is a (possibly infinite) set of formulas with only finitely many distinct variables occuring in it, then $\bigwedge\Phi$ is a formula of L^*. It is easy to verify that for every sentence ϕ of L, ϕ^* is a sentence of L^*.

An assignment σ for L^* is an ordered pair $\langle\sigma_I, \sigma_P\rangle$, where σ_I is an Ω-sequence of finitely many individuals, and σ_P analogously for properties. An *L^*-isomorphism* between world-assignment pairs is defined as follows: $\langle w,\sigma_I,\sigma_P\rangle$ is L^*-isomorphic to $\langle w',\sigma_I',\sigma_P'\rangle$ iff all of the following conditions hold: (i) Actuality-Preservation: (a) If $w = @$ or $w' = @$, then $w = w'$. (b) If $\sigma_I(n) = a$ or $\sigma_I'(n) = a$, then $\sigma_I(n) = \sigma_I'(n)$. (c) If $\sigma_P(n) = P$ or $\sigma_P'(n) = P$, then $\sigma_P(n) = \sigma_P'(n)$. (ii) Identity-Preservation: (a) $\sigma_I(n) = \sigma_I(m)$ iff $\sigma_I'(n) = \sigma_I'(m)$. (b) Likewise for the property assignment. (iii) Instantiation-Preservation: $\sigma_I(n)$ and $\sigma_P(m)$ are worldmates iff $\sigma'_I(n)$ and $\sigma'_P(m)$ are worldmates. (iv) Existence-Preservation: (a) $\sigma_I(n)$ is in w iff $\sigma'_I(n)$ is in w'. (b) $\sigma_P(n)$ is in w iff $\sigma'_P(n)$ is in w'.

---

[28] For the idea to use that mapping and Lemma 1 in the proof, I am indebted to Burgess.

> *Lemma 2.* For any formula ϕ of L^*, worlds w and w' and assignments σ and σ' such that $\langle w, \sigma\rangle$ and $\langle w', \sigma'\rangle$ are isomorphic, $\mathcal{M}_{w\sigma} \models \phi$ iff $\mathcal{M}_{w'\sigma'} \models \phi$.

Proof. It is routine to check that the conditons on L^*-isomorphisms ensure that this holds for atomic formulas. The inductive clauses for negation and conjunction are straightforward. Actuality operator: $\mathcal{M}_{w\sigma} \models A\phi$ iff $\mathcal{M}_{@\sigma} \models \phi$. Since $\langle @,\sigma\rangle$ and $\langle @,\sigma'\rangle$ are isomorphic, $\mathcal{M}_{@\sigma'} \models \phi$ by the induction hypothesis. Hence $\mathcal{M}_{w\sigma'} \models A\phi$. Diamond: suppose $\mathcal{M}_{w\sigma} \models \Diamond\phi$. By definition, $\mathcal{M}_{w^*\sigma} \models \phi$ for some world w^*. Then there is a world $w^{*\prime}$ such that $\langle w^*,\sigma\rangle$ and $\langle w^{*\prime},\sigma'\rangle$ are isomorphic. Thus $\mathcal{M}_{w^{*\prime}\sigma'} \models \phi$ by the induction hypothesis. Hence $\mathcal{M}_{w'\sigma'} \models \Diamond\phi$. Existential quantification: suppose $\mathcal{M}_{w\sigma} \models \exists x\, \phi(x)$. Then $\mathcal{M}_{w\sigma(x/w)} \models \phi(x)$. Since $\langle w,\sigma(x/w)\rangle$ and $\langle w',\sigma'(x/w')\rangle$ are isomorphic, $\mathcal{M}_{w'\sigma'(w/x)} \models \phi(x)$ by the induction hypothesis, and thus $\mathcal{M}_{w'\sigma'} \models \exists x\ \phi(x)$. The clause for the property quantifier is exactly analagous.

A *shared* world is one that is in both $\mathcal{M}$ and $\mathcal{M}'$; since the latter is a subset of the former, these are just the worlds in $\mathcal{M}'$. A shared assignment is one that only assigns individuals and properties from shared worlds.

> *Lemma 3.* For any formula ϕ of L^*, shared assignment σ and shared world w, $\mathcal{M}_{w\sigma} \models \phi$ iff $\mathcal{M}'_{w\sigma} \models \phi$.

Proof. For atomic formulas, this is immediate from the construction of the models. The clauses for negation, conjunction, and the actuality-operator are straightforward. Suppose $\mathcal{M}_{w\sigma} \models \Diamond\phi$. Then there is a world w' such that $\mathcal{M}_{w'\sigma} \models \phi$. For some shared world w'', $\langle w',\sigma\rangle$ and $\langle w'',\sigma\rangle$ are isomorphic. (Either w' is shared, in which case $w' = w''$; or w' is not shared, then since σ is shared, w'' may be any shared world not involved in σ.) By Lemma 2 and the induction hypothesis, $\mathcal{M}'_{w''\sigma} \models \phi$. Hence $\mathcal{M}'_{w\sigma} \models \Diamond\phi$. The other direction is straightforward. For existential quantification, suppose $\mathcal{M}_{w\sigma} \models \exists x\ \phi(x)$. Then $\mathcal{M}_{w\sigma(x/w)} \models \phi(x)$. Since $\sigma(x/w)$ is obviously a shared assignment, it follows that $\mathcal{M}'_{w\sigma(x/w)} \models \phi(x)$ by the induction hypothesis, and hence $\mathcal{M}'_{w\sigma} \models \exists x\, \phi(x)$. Again, the other direction is straightforward. The clause for the property quantifier is exactly analogous.

> *Proposition.* The non-isomorphic modal models $\mathcal{M}$ and $\mathcal{M}'$ are elementarily equivalent relative to L; that is for every sentence ϕ of L, $\mathcal{M} \models \phi$ iff $\mathcal{M}' \models \phi$.

Proof. The harder direction is right-left: suppose $\mathcal{M}' \models \phi$, that is $\mathcal{M}'_{@\sigma} \models \phi$ for every $\mathcal{M}'$-assignment (shared assignment) σ. By Lemma (1.b), $\mathcal{M}'_{@\sigma} \models \phi^*$ for every shared σ. By Lemma 3, and since @ is a shared world, $\mathcal{M}_{@\sigma} \models \phi^*$ for every shared assignment. By Lemma 2, and since for every σ, there is some shared σ' such that $\langle @,\sigma \rangle$ and $\langle @,\sigma' \rangle$ are isomorphic, $\mathcal{M}_{@\sigma} \models \phi^*$ for every σ. By Lemma (1.a), $\mathcal{M}_{@\sigma} \models \phi$.

THE JOURNAL OF
PHILOSOPHY

BACK ISSUE SALE

Complete volumes
are now available
at the following discounted prices
in order to reduce our inventory:

Volumes	1 through 98	$10 each
Volumes	99 through 101	$25 each

For more information, contact Pamela Ward,
Business Manager, at pcw1@columbia.edu

Visa, Mastercard, American Express and
PayPal are accepted through our website at
www.journalofphilosophy.org

the review of

metaphysics

a philosophical quarterly

JUNE 2006 | **VOL. LI X, No. 4** | **ISSUE No. 236** | **$15.00**

articles

SARAH BYERS *Life as "Self-Motion": Descartes and the Aristotelians on the Soul as the Life of the Body*

RICHARD DIEN WINFIELD *Self-Consciousness and Intersubjectivity*

KENNETH WESTPHAL *How Does Kant Prove that We Perceive, and Not Merely Imagine, Physical Objects?*

DALIA NASSAR *Immediacy and Mediacy in Schleiermacher's* Reden Uber die Religion

DOUGLAS B. RASMUSSEN and DOUGLAS DEN UYL *The Myth of Atomism*

ELIZABETH C. SHAW AND STAFF *Summaries and Comments*

book reviews *philosophical abstracts* *announcements*

Individual Subscriptions $35.00 Institutional Subscriptions $55.00 Student/Retired Subscriptions $20.00

VISA AND MASTERCARD ACCEPTED

Jude P. Dougherty, Editor
The Catholic University of America, Washington D.C. 20064
email: mail@reviewofmetaphysics.org

THE

Philosophical Review

Duke University Press
is pleased to be the new publisher of
the *Philosophical Review* beginning in 2006.

> *"No philosophy journal published in English is more highly regarded than the* Philosophical Review."
> *—David Sanford, Duke University*

In continuous publication since 1892, the *Philosophical Review* has a long-standing reputation for excellence and has published many papers now considered classics in the field, such as W. V. O. Quine's "Two Dogmas of Empiricism," Thomas Nagel's "What Is It Like to Be a Bat?" and the early essays of John Rawls. The journal aims to publish original scholarly work in all areas of analytic philosophy, with an emphasis on material of general interest to academic philosophers, and is one of the few journals in the discipline to publish book reviews.

Subscription Information
Quarterly
Individuals: $33
Students: $20 (photocopy of student ID required)
Single issues: $12

Edited by the faculty of the Sage School of Philosophy at Cornell University

Subscribers outside the U.S. and Canada:
Please add $16 for postage.
Canadian subscribers:
Please add 7% GST and $12 for postage.

To place your order using a credit card, please call toll-free 888-651-0122 (in the U.S. and Canada) or 919-688-5134, or e-mail **subscriptions@dukeupress.edu.**

For more information, please visit **www.dukeupress.edu/philreview.**

DUKE
UNIVERSITY PRESS

Ethics
an international journal of
social political and legal philosophy

John Deigh, *Editor*

Founded in 1890, ***Ethics*** publishes scholarly work in moral, political, and legal philosophy from a variety of intellectual perspectives, including social and political theory, law, and economics. In addition to major atricles, ***Ethics*** also publishes review essays, discussion articles, book reviews, and book notes.

Quarterly. ISSN: 0014-1704

***Ethics* Online**

Read a sample issue, sign up for RSS feeds, and register for e-TOC tables of contents alerts at ***www.journals.uchicago.edu/ET***

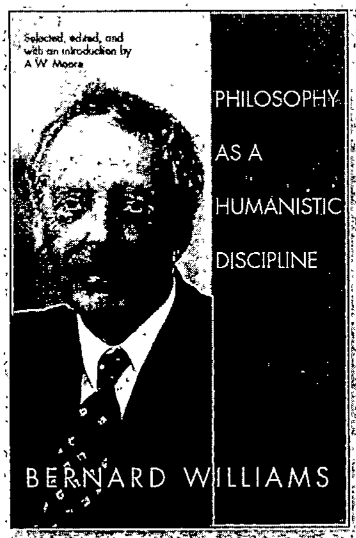
Subscribe by:

Phone: Toll-free at (877) 705-1878 (US/Canada), or (773) 753-3347 (International)
Fax: Toll-free at (877) 705-1879 (US/Canada), or (773) 753-0811 (International)
Online: *www.journals.uchicago.edu/ET*
E-mail: *subscriptions@press.uchicago.edu*

The University of Chicago Press
Journals Division • P.O. Box 37005 • Chicago, IL 60637
www.journals.uchicago.edu

3/06

Original


Selected, edited, and with an introduction by A W Moore
PHILOSOPHY AS A HUMANISTIC DISCIPLINE
BERNARD WILLIAMS

Philosophy as a Humanistic Discipline

Selected, edited, and with an introduction by A. W. Moore

BERNARD WILLIAMS

In *Philosophy as a Humanistic Discipline*, Bernard Williams presents a striking vision of philosophy as fundamentally different from science in its aims and methods. Written with his distinctive combination of rigor, imagination, depth, and humanism, the book amply demonstrates why Williams was one of the greatest philosophers of the twentieth century.

"This is a splendid collection."
—Richard Moran, author of *Authority and Estrangement*

Cloth $35.00 0-691-12426-4

From the author of the New York Times *Bestseller,* On Bullshit

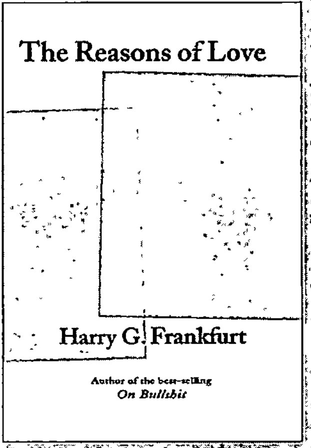
The Reasons of Love

HARRY G. FRANKFURT

"A pleasure to read. . . . [This book] is a comprehensive statement of the mature views of one of the most creative philosophers of his generation."
—Philip L. Quinn,
Notre Dame Philosophical Reviews

Winner of the 2004 Award for Best Professional/Scholarly Book in Philosophy, Association of American Publishers

New in paperback $12.95 0-691-12624-0



100 Years of Excellence

PRINCETON
University Press

800-777-4726
Read excerpts online
www.pup.princeton.edu

CAMBRIDGE

BEST IN SCHOLARSHIP

Kant: Anthropology from a Pragmatic Point of View
Edited by Robert B. Louden and Manfred Kuehn
Cambridge Texts in the History of Philosophy
$65.00: Hardback: 0-521-85556-X: 288pp
$22.99: Paperback: 0-521-67165-5

Knowledge and Inquiry
Erik J. Olsson
Cambridge Studies in Probability, Induction and Decision Theory
$80.00: Hardback: 0-521-84556-4: 408pp

Moral Disagreement
Folke Tersman
Cambridge Studies in Philosophy
$60.00: Hardback: 0-521-85338-9: 160pp

The New Hegelians
Douglas Moggach
$75.00: Hardback: 0-521-85497-0: 360pp

A Theory of Argument
Mark Vorobej
$75.00: Hardback: 0-521-85431-8: 334pp
$29.99: Paperback: 0-521-67025-X

Philosophical Perspectives on Infinity
Graham Oppy
$75.00: Hardback: 0-521-86067-9: 336pp

Nietzsche's Philosophy of Religion
Julian Young
$75.00: Hardback: 0-521-85422-9: 217pp
$29.99: Paperback: 0-521-68104-9

Science and Ethics
Bernard Rollin
$70.00: Hardback: 0-521-85754-6: 304pp
$25.99: Paperback: 0-521-85790-2

Nietzsche: Human, All Too Human
A Book for Free Spirits
Friedrich Nietzsche
Edited by R. J. Hollingdale and Richard Schacht
Cambridge Texts in the History of Philosophy
$18.99: Paperback: 0-521-56704-1: 430pp

Nietzsche: The Gay Science
With a Prelude in German Rhymes and an Appendix of Songs
Friedrich Nietzsche
Edited by Bernard Williams, Josefine Nauckhoff, and Adrian Del Caro
Cambridge Texts in the History of Philosophy
$12.99: Paperback: 0-521-63645-0: 308pp

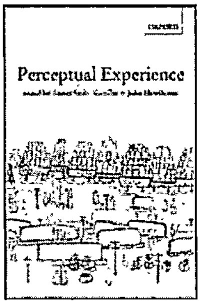
Nietzsche: Beyond Good and Evil
Friedrich Nietzsche
Edited by Rolf-Peter Horstmann and Judith Norman
Cambridge Texts in the History of Philosophy
$14.99: Paperback: 0-521-77913-8: 230pp

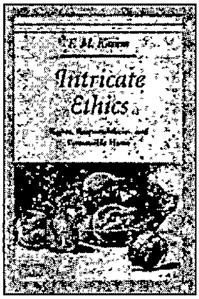
The Principle of Sufficient Reason
A Reassessment
Alexander Pruss
$80.00: Hardback: 0-521-85959-X: 352pp

Prices subject to change.

www.cambridge.org/us/philosophy/

CAMBRIDGE
UNIVERSITY PRESS





New from
Oxford

Perceptual Experience

Edited by TAMAR SZABO GENDLER
and JOHN HAWTHORNE

Perceptual Experience is comprised of papers written specially for this volume, and they cover a broad range of topics dealing with sensation and representation, consciousness and awareness, and the connections between perception and knowledge and between perception and action. Contributors include John Campbell, David J. Chalmers, Tim Crane, Fred Dretske, Tamar Szabo Gendler, Anil Gupta, John Hawthorne, Susan Hurley, Mark Johnston, Geoffrey Lee, Eric Lormand, M. G. F. Martin, Alva Noë, Jesse J. Prinz, Sydney Shoemaker, Susanna Siegel, and Michael Tye.

February 2006 paper $35.00 cloth $99.00

Intricate Ethics

Rights, Responsibilities, and Permissible Harm

F. M. KAMM

In *Intricate Ethics*, Kamm questions the moral importance of some non-consequentialist distinctions and then introduces and argues for the moral importance of other distinctions. The first section provides a general introduction to non-consequentialist ethical theory; the second deals with the notions of moral status and rights; the third takes up the notions of responsibility and complicity; and the fourth section analyzes the views of others in the non-consequentialist and consequentialist camps.

March 2006 $65 00

Analyzing Oppression

ANN E. CUDD

Analyzing Oppression presents a new, integrated theory of social oppression, which tackles the fundamental question that no theory of oppression has satisfactorily answered: if there is no natural hierarchy among humans, why are some cases of oppression so persistent? Cudd argues that the explanation lies in the coercive co-opting of the oppressed to join in their own oppression.

March 2006 paper $35 00 cloth $74 00

My Way

Essays on Moral Responsibility

JOHN MARTIN FISCHER

This is a selection of essays on moral responsibility that represent the major components of John Martin Fischer's overall approach to freedom of the will and moral responsibility.

February 2006 $45.00

Choosing Children

The Ethical Dilemmas of Genetic Intervention

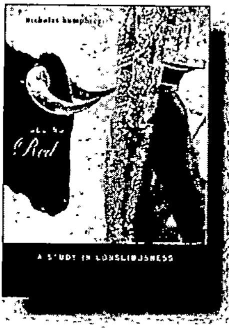
JONATHAN GLOVER

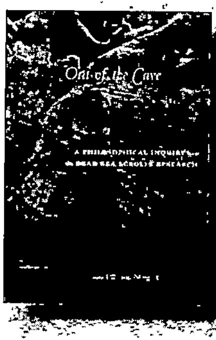
Genetic and reproductive technology now offers us the possibility of choosing what kinds of children we do and don't have. Should we welcome this power, or should we fear its implications? The renowned moral philosopher and best-selling author Jonathan Glover suggests how we should try to deal with this urgent problem.

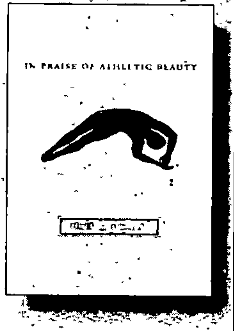
March 2006 $15.95

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS







SEEING RED

A Study in Consciousness

NICHOLAS HUMPHREY

Beginning with the seemingly simple act of seeing red, the fact of a conscious sensation, this brilliantly unsettling essay by one of our most interesting writers on human intelligence builds toward an explanation of why consciousness matters—why, in short, it makes compelling evolutionary sense.

Belknap Press • new in cloth

OUT OF THE CAVE

A Philosophical Inquiry into the Dead Sea Scrolls Research

EDNA ULLMANN-MARGALIT

Looking at the Essene connection, the archaeology of Qumran, and the sectarian nature of the scrolls community, Ullmann-Margalit explores the different arenas, and ways, in which contesting theories of the scrolls do battle.

new in cloth

IN PRAISE OF ATHLETIC BEAUTY

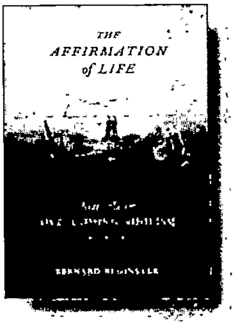
HANS ULRICH GUMBRECHT

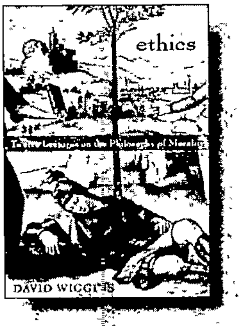
"With this study of athletic aesthetics, Stanford literature professor Gumbrecht scores as the sports world's answer to Marshall McLuhan...Gumbrecht's writing is as potent and graceful as the athletes he admires."

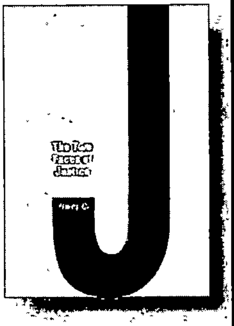
—PUBLISHERS WEEKLY

Belknap Press • new in cloth

WWW.HUP.HARVARD.EDU HARVARD UNIVERSITY PRESS







THE AFFIRMATION OF LIFE

Nietzsche on Overcoming Nihilism

BERNARD REGINSTER

"With lucid writing and a superb examination of his works, Reginster succeeds in presenting a coherent interpretation of Nietzsche's philosophy.
—Scott Duimstra, LIBRARY JOURNAL
new in cloth

ETHICS

Twelve Lectures on the Philosophy of Morality

DAVID WIGGINS

Almost every thoughtful person wonders at some time why morality says what it says and how, if at all, it speaks to us. David Wiggins surveys the answers most commonly proposed for such questions.
new in cloth

THE TWO FACES OF JUSTICE

JIWEI CI

Rather than pursue a search for normative principles, Jiwei Ci probes the human psychology of justice to understand what motivates moral agents who seek to behave justly, and why their desire to be just is as precarious as it is uplifting.
new in cloth

WWW.HUP.HARVARD.EDU HARVARD UNIVERSITY PRESS

COMING SOON:

THE JOURNAL OF PHILOSOPHY

ONLINE

The Journal of Philosophy will be adding
a print and online subscription option for institutions
to begin Fall 2006

SUBSCRIPTION OPTIONS

Libraries and Institutions (print and online)	$180.00
Libraries and Institutions (print only)	$100.00
Individuals	$45.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

We expect to launch this new feature with our September 2006 issue.
Please consult our website at www.journalofphilosophy.org
for more details.

Calcutta University
Batta 9/11/06
ALIPORE
CAMPUS

THE JOURNAL OF PHILOSOPHY

VOLUME CIII, NUMBER 4
APRIL 2006

page

163 *Libertarianism and the Philosophical Significance of Frankfurt Scenarios* David Widerker
188 *Control, Choice, and the Convergence/Divergence Dynamics: A Compatibilistic Probabilistic Theory of Free Will* Marius Usher
214 NEW BOOKS: REPRINTS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Christopher Peacocke, Carol Rovane, Achille C. Varzi. *Consulting Editors*: David Albert, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2006

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2006 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIII, NO. 4, APRIL 2006

LIBERTARIANISM AND THE PHILOSOPHICAL SIGNIFICANCE OF FRANKFURT SCENARIOS*

It seems uncontroversial that

> *Freedom Condition* (FC) An agent S is morally blameworthy for a performing a given act V only if S performed V freely.

But how exactly is "freely" in FC to be understood? The traditional answer to this question has been to identify the freedom pertinent to moral blame with avoidability, thus construing FC as follows:

> *Principle of Alternate Possibilities* (PAP) An agent S is morally blameworthy for performing a given act V only if he could have avoided performing it.[1]

---

*I would like to thank Paul Bartha, Ori Beck, Bob Bunn, Steve Darwall, Jerome Gellman, Dovid Gottlieb, Bernard Katz, Jonathan Schaffer, Angela Smith, and most especially Michael Feld, Carl Ginet, Bill Rowe, Ira Schnall, and an editor of this JOURNAL for their excellent comments and discussions on earlier versions of this paper. An earlier version of this paper was also presented at the departmental research seminar at Bar-Ilan University in November 2005, and at the Inland Northwest Philosophy Conference on Action, Ethics, and Responsibility in March 2006. I would like to thank the audiences at those presentations for their helpful comments.

[1] I am assuming a fine-grained account of action individuation according to which an action is a dated particular consisting at least in part in an agent's exemplifying an act-property at a time. See Alvin Goldman, *A Theory of Human Action* (Englewood Cliffs, NJ: Prentice Hall, 1971), pp. 15–17.

The term "blameworthy" in FC and PAP refers only to *direct or nonderivative* culpability, as opposed to *indirect* or *derivative* culpability. An agent is *directly* or *nonderivatively* blameworthy for performing an act V only if he is blameworthy for doing so, but not in virtue of being blameworthy for some other act or fact. Otherwise he is *indirectly* or *derivatively* blameworthy for doing V. A typical case of derivative culpability is a scenario in which an agent, who is aware that doing V at T is morally wrong, deliberately places himself in circumstances where he loses his power to avoid doing V at T. If ultimately, the agent does V at T, we say that he is *derivatively* blameworthy for doing V at T, even though (shortly before T) he could not have avoided doing so.

0022-362X/06/0304/163–87

Situations in which the agent is coerced, compelled, or manipulated to perform a certain act confirm PAP. As is well known, PAP has come under attack most notably by Harry Frankfurt. In his seminal article "Alternate Possibilities and Moral Responsibility," Frankfurt develops an intriguing argument against PAP.[2] Central to it are two assumptions: one metaphysical and one moral. The metaphysical assumption is that

> (IRR) There may be circumstances that in no way bring it about that a person performs a certain action; nevertheless, those very circumstances make it impossible for him to avoid performing that action (*ibid.*, pp. 830, 837).

(Call a situation of the sort described in IRR an "IRR- scenario.") The moral assumption is that an agent who in an IRR-situation acted wrongly is blameworthy for what he did. Together these two assumptions yield the conclusion that PAP is false. We may get a better grasp of Frankfurt's argument by considering the kind of example to which Frankfurt appeals in order to establish IRR.

> Jones is deliberating as to whether to keep the promise he made to his uncle to visit him in the hospital shortly before a critical operation his uncle is about to undergo. Jones is his uncle's only relative, and the visit is very important to the uncle. The reason for Jones's deliberating is that, on his way to the hospital, he (incidentally) met Mary—a woman with whom he was romantically involved in his distant past and whom he has not seen since then. Mary, being eager to talk to Jones, invites him for a cup of coffee in a nearby restaurant. She explains that she is in town just for a couple of hours, and wishes to spend those hours with him. Jones is aware that if he accepts Mary's offer, he will not be able to make it to the hospital during visiting hours. Unbeknownst to Jones, there is another person, Black, who for some reason does not want Jones to visit his uncle. Black has the power and the means to force Jones to decide to stay with Mary. But wishing to avoid showing his hand unnecessarily, he has made up his mind to intervene if and only if Jones does *not* show a sign that he is going to decide to break his promise to his uncle. Call that sign *Q*. If Jones shows that sign, then Black does nothing, knowing that in this case Jones will decide to accept Mary's invitation. (It is assumed that Black knows Jones very well in this regard.) Finally, suppose that Black does not have to intervene, since Jones decides on his own not to keep the promise, so as to be able to spend time with Mary.

---

[2] Frankfurt, "Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 4, 1969): 829–39; reprinted in David Widerker and Michael McKenna, eds., *Moral Responsibility and Alternative Possibilities* (Aldershot, UK: Ashgate, 2003), pp. 17–25 (page references here will be to the original).

Call this example "Promise Breaking" and an example of this type a "Frankfurt-type counterexample to PAP," or a "Frankfurt-type example" for short. Frankfurt claims that, because in the situation described Jones acted freely (without Black's intervention), he is to blame for deciding to break the promise, even though he could not have avoided doing so. Hence, according to him, PAP is false (*ibid.*, pp. 837–88).

Frankfurt's attack on PAP has not gone unchallenged. Libertarians have resisted it by questioning the coherence of IRR-situations, or at any rate by insisting that examples such as the above fail to describe an IRR-situation, in which case, Frankfurt's assumption IRR is unwarranted. More specifically, they have claimed that such examples do not succeed when applied to mental acts such as deciding, undertaking, forming an intention, that is, mental acts that for the libertarian constitute the *loci* of moral responsibility. Their counterargument (sometimes also referred to as the "Dilemma Objection") was this:

> Consider the relation between the sign Q that Black employs as a sign for not intervening and Jones's decision not to keep the promise. (Call that decision $D(-K)$.) Either the occurrence of Q is (or is indicative of) a condition that, in the circumstances, is causally sufficient for $D(-K)$, or it is not. If it is, then $D(-K)$ is brought about by the circumstances in which it occurs, in which case the example does not describe an IRR-situation. On the other hand, if Q is not so associated with $D(-K)$, but is merely a reliable indicator of it, then there is no reason to think that Jones's decision was unavoidable. In either case, the truth of IRR has not been established.[3]

To sidestep this libertarian response, opponents of PAP have tried to come up with other, more sophisticated, examples of IRR-scenarios.[4] These examples have again been criticized by PAP-defenders,[5] and then have again been improved upon or defended by opponents of PAP,[6] and so on.

---

[3] For an elaborate version of this argument, see Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *Philosophical Review*, CIV (1995): 247–61. For similar responses to Frankfurt's argument against PAP, see Carl Ginet, "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing," *Philosophical Perspectives*, X (1996): 403–17; Robert Kane, *Free Will and Values* (Albany: SUNY, 1985), p. 51; Kane, *The Significance of Free Will* (New York: Oxford, 1996), pp. 142–43; James W. Lamb, "Evaluative Compatibilism and the Principle of Alternate Possibilities," this JOURNAL, XC, 10 (October 1993): 497–516. For a libertarian response to Frankfurt of a Reidian type, see William L. Rowe, *Thomas Reid on Freedom and Morality* (Ithaca: Cornell, 1991), pp. 82–85.

[4] See, for example, David Hunt, "Frankfurt Counterexamples: Some Comments on the Widerker-Fischer Debate," *Faith and Philosophy*, XIII (1996): 395–401; David Hunt, "Moral Responsibility and Unavoidable Action," *Philosophical Studies*, XCVII (2000): 195–227; Alfred Mele and David Robb, "Rescuing Frankfurt-Style Cases," *Philosophical Review*, CVII (1998): 97–111; Eleonore Stump, "Moral Responsibility Alternative

The worry behind the Dilemma Objection is understandable. If in an IRR-situation the agent's decision is not causally determined, then why is it not within the agent's power to refrain from it? In the past, I have been skeptical about the prospects of providing a successful example of an IRR-situation. Since then, however, I have come to believe that such situations *are* conceptually possible. My purpose in this paper is to explain why this is so, and, as a result, to offer a novel perspective on the significance of such situations for a *libertarian* account of freedom and moral blameworthiness.

The discussion that follows is divided into five parts. In sections I–II, I present my argument for the conceptual possibility of IRR-situations. In section III, I apply my account to Frankfurt-type-examples. In this context, I criticize Derk Pereboom's recent attempt at providing an adequate Frankfurt-type example, and draw attention to a general difficulty that threatens their cogency. In section IV, I show how the notion of an IRR-situation can be further generalized. Section V is a dialectical interlude in which I elaborate on the philosophical significance of IRR-scenarios. In sections VI–VII, I develop what I call an "avoidability-independent account of moral blame" which is compatible with the rejection of PAP. And finally, in section VIII, I defend that account against some potential objections.

I. HOW TO CONSTRUCT A SUCCESSFUL EXAMPLE OF AN IRR-SITUATION

I begin with a few definitions:

(D1) By *state of affairs* I am referring to dated states of affairs consisting in an object exemplifying a property at a certain time. For example, that Nixon was born on January 9, 1913; that Columbus discovers Cuba on October 28, 1492, and so on.

Possibilities: The Flicker of Freedom," *The Journal of Ethics*, III (1999): 299–324; and Derk Pereboom, "Alternative Possibilities and Causal Histories," *Philosophical Perspectives*, XIV (2000): 119–37.

[5] See, for example, Widerker, "Frankfurt's Attack on the Principle of Alternative Possibilities: A Further Look," *Philosophical Perspectives*, XIV (2000): 181–201; Carl Ginet's addendum to his "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing," in Widerker and McKenna, eds., pp. 85–87; Robert Kane, "Responsibility, Indeterminism, and Frankfurt- Style Cases," in Widerker and McKenna, eds., pp. 91–105; Stuart Goetz, "Stump on Libertarianism and Alternative Possibilities," *Faith and Philosophy*, XVIII (2001): 93–101; Stuart Goetz, "Alternative Frankfurt-Style Counterexamples to the Principle of Alternative Possibilities," *Pacific Philosophical Quarterly*, LXXXIII (2002): 131–47.

[6] See, for example, Mele and Robb, "Bbs, Magnets and Seesaws: The Metaphysics of Frankfurt-Style Cases," in Widerker and McKenna, eds., pp. 127–38; Stump, "Moral

(D2) A state of affairs E is *causally possible,* relative to a time $T =_{\text{def}}$ E obtains in some possible world W that up until (and including) T has the same history and the same laws of nature as the actual world.

(D3) An actual state of affairs E is *nomically determined* by a temporally prior state of affairs $F =_{\text{def}}$ E is entailed by the conjunction of F and some causal laws, and is not entailed by either conjunct alone.

(D4) An actual state of affairs E is *nomically determined* (simpliciter) $=_{\text{def}}$ E is nomically determined by some temporally prior state of affairs.

(D5) An actual state of affairs F indexed to a time T *deterministically causes* a later state of affairs $E =_{\text{def}}$ F causes E, and E obtains in every possible world W that up until (and including) T has exactly the same history and the same causal laws as the actual world.

(D6) An actual state of affairs F indexed to a time T *indeterministically causes* a later state of affairs $E =_{\text{def}}$ F causes E, and it is causally possible (relative to T) that F does not cause E.

(D7) An actual state of affairs E is *causally determined* $=_{\text{def}}$ E is either nomically determined or is deterministically caused by some prior state of affairs.

(D8) An actual state of affairs E indexed to a time T is *causally undetermined* $=_{\text{def}}$ E is not causally determined (which implies that there is a possible world W that up to T (but not including T) has the same history and the same causal laws as the actual world, in which E does not obtain).

At the heart of my argument for the conceptual possibility of an IRR-situation are two assumptions. The first concerns a distinction familiar from discussions of free will; the distinction between its being *causally possible* (relative to a time T) that a certain state of affairs obtains, and an agent's *having it within his power* at T to bring it about that the state of affairs obtains. Obviously, the fact that a state of affairs is causally possible, relative to T, say, that a radioactive atom will decay at some later time T', does not entail that an agent has it within his power at T to bring that state of affairs about.

Call a state of affairs (event) E *actionally accessible* for an agent S, relative to a time T, if and only if S has it within his power at T to bring about E. Call a state of affairs (event) E *actionally inaccessible* for an agent S, relative to a time T, if and only if S does *not* have it within his power at T to bring about E. These two kinds of states of affairs (hereafter also referred to simply as "*accessible/inaccessible* states of affairs") may be either actual or nonactual. Of special interest to us

---

Responsibility without Alternative Possibilities," in Widerker and McKenna, eds., pp. 139–58; Hunt, "Moral Responsibility and Buffered Alternatives," *Midwest Studies in Philosophy,* XXIX (2005): 126–45.

will be *inaccessible* states of affairs that are not actual. The notion of an *inaccessible/accessible* state of affairs is both agent and time-relative. For example, normally, the causally possible (nonactual) state of affairs of Carl's inviting me for dinner at T_1 is inaccessible for me before T_1, but not for him. However, after T_1, that state of affairs becomes inaccessible for him as well. Also, some states of affairs (events) are *inaccessible* for every human—for example, that the universe ceases to exist in January 1800.

The second assumption in my argument for the possibility of an IRR-situation is that it is a situation in which an agent S makes a certain decision *on his own*, without having a (for him) accessible alternative to that decision. The term "on his own" is to be understood in a libertarian sense, implying that such a decision is neither caused[7] nor nomically determined.[8]

Expressing all this a bit more formally, we can say that an IRR-situation is a situation in which:

(a) S decides-to-V at T on his own.[9]
(b) $P_1, P_2, \ldots, P_n$ are *all* (in the circumstances) causally possible alternatives to S's deciding-to-V at T.
(c) Each P_i ($i = 1– n$) is *inaccessible* to S in the circumstances.[10]

Letting "$D^O(V, S, T)$" to stand for "S decides-to-V at T *on his own*," we may usefully characterize an IRR-situation as a situation in which the following disjunction holds:

(IRR*) $D^O(V, S, T) \vee P_1 \vee P_2 \vee \ldots \vee P_n$,

where $D^O(V, S, T)$ is true in that situation, and $P_1, P_2, \ldots, P_n$ are as before.

---

[7] By "caused" I mean either deterministically or indeterministically caused. There are, of course, various accounts of indeterministic causation. Here I am focusing on an account according to which the obtaining of the causal relation is not within the agent's control, that is, that if E indeterministically causes F, then once E obtains, its causing F is not something that is up to the agent. On this point, see Peter van Inwagen, *An Essay on Free Will* (New York: Oxford, 1983), p. 144, and Timothy O'Connor, *Persons and Causes* (New York: Oxford, 2000), p. 29.

[8] This notion of "acting on one's own" is closely related to Kane's notion of a *self-forming* act, or to Pereboom's notion of agent's being the *ultimate source* of his act. See Kane, *The Significance of Free Will*, pp. 124–25, and Pereboom, *Living without Free Will* (New York: Cambridge, 2001), pp. 4–5. It differs from Kane's notion, since for Kane a libertarianly free act may be indeterministically caused.

[9] The T in "S decides-to-V at T" refers to the time at which the decision is made, and not to the time at which the act of V-ing is performed.

[10] These circumstances extend from the onset of the IRR-situation (= the time at which the agent's the deliberation process begins) until T. Typically, the reason why a given P_i is inaccessible to S, is either that S does not know how to bring that P_i about, or that (in the circumstances) S's having the power to realize that P_i is contingent upon the obtaining of a state of affairs (event) that is inaccessible to him in the cir-

This characterization of an IRR-situation is in line with its initial definition (p. 164) as a situation in which the agent's decision, though unavoidable, is in no way brought about by the circumstances in which it occurs. The task that remains is to present a convincing case for the claim that such situations are conceptually possible. It is this task to which I now turn.

II. TWO EXAMPLES OF AN IRR-SITUATION

For expository purposes, let me begin with a somewhat unusual example.

1. Z-Persons. Suppose that neuroscientists were to discover that persons of a certain type Z behave in the following way: when in a deliberative situation in which they are strongly tempted to act immorally, they invariably either succumb to the temptation and make the wrong decision, or they lose consciousness for one second, and regain it thereafter. Now suppose that Jones is such a person, and he is in the kind situation described in Promise Breaking (p. 164), except that this time there is no Black waiting in the wings. Suppose also that Jones does not know that he is subject to the said behavioral pattern. As things transpire, Jones decides at time T to break the promise on his own.

I claim that the situation before us is an IRR-situation. To see this, consider the disjunctive fact:

(1) Either Jones decides at T on his own not to keep the promise, or Jones faints at T.

Given that Jones is a Z-person, this fact is nomically determined, which is consistent with Jones's decision not to keep the promise ($D(-K)$) being neither nomically determined nor caused, that is, consistent with its being a decision that Jones makes on his own. On the other hand, it is not within Jones's power to avoid making that decision, since the only way in which he could do so is by becoming unconscious at T, which state of affairs is *actionally inaccessible* to him. (Here I am assuming that becoming unconscious is not something that is within Jones's voluntary control.)

2. Brain-Malfunction-W.[11] This example also does not involve a counterfactual intervener. As in Promise Breaking, Jones deliberates as to whether to accept Mary's offer, and ultimately decides on his

---

cumstances. A more economical (yet still equivalent) way of defining an IRR-situation would be to say that it is a situation in which S decides-to-V at T on his own, such that the negative state of affairs *that it is not the case that S decides-to-V at T* can be realized only by states of affairs that, in the circumstances, are inaccessible to S.

[11] The "*W*" in "Brain-Malfunction-*W*" indicates that I wish to distinguish it from other examples of that sort, specifically from the one given by McKenna, "Robustness,

own at *T* to violate the promise to his uncle. Normally, one can avoid deciding as one does by deciding otherwise. But in our scenario Jones does not have that option, since shortly before beginning to deliberate, he undergoes a neurological change as a result of which one of the (neurological) causally necessary conditions for his deciding otherwise, a condition which we may call *N*, does not obtain.[12] Let us also assume that all this is unknown to Jones (who believes that he can decide to keep the promise), and that *N*'s absence does not affect his deliberation process.[13] Note that these assumptions do not render Jones's actual decision of not keeping the promise ($D(-K)$) causally determined. The fact that *N* is a causally necessary condition of Jones's deciding to keep the promise entails that *N*'s absence is sufficient for its *not* being the case that Jones decides to keep it, that is, for $-D(K)$. But from this it does not follow that *N*'s absence is sufficient for $D(-K)$. $D(-K)$ and the negative state of affairs $-D(K)$, are not identical. Though in the situation under consideration $-D(\mathrm{K})$ is realized by $D(-K)$, it can be also realized in a different way. For example, by Jones's suddenly becoming distracted and forgetting for a few seconds about the decision he needs to make, or by Jones's spontaneously losing focus for a second or two. Thus, the absence of *N* does not causally determine $D(-K)$.

We are now very close to obtaining an IRR-situation. I say "very close," as we have not yet eliminated *all* the alternatives to Jones's deciding not to keep the promise that are *accessible* to him. There is the alternative of his continuing to deliberate at *T*, instead of making $D(-K)$. However, this alternative can be ruled out either by stipulation or by assuming that in the scenario at hand Jones needs to make a decision right away, as otherwise he will miss the bus that can get him to the hospital on time. If that is the case, then in our scenario, the option of continuing to deliberate is practically equivalent to deciding not to keep the promise.

Control, and the Demand for Morally Significant Alternatives," in Widerker and McKenna, eds., pp. 209–10. (This example is based on an earlier version of Brain-Malfunction-*W*.) For a brief, decisive, criticism of McKenna's example, see Widerker, "Blameworthiness, Non-Robust Alternatives, and the Principle of Alternative Expectations," *Midwest Studies in Philosophy*, XXIX (2005): 292–306, footnote 15.

[12] Note also that *N* need not be a neurological condition. Dualists may assume that *N* is a mental event/state of some sort.

[13] I am assuming that *N* is not only a necessary condition for Jones's deciding otherwise at *T*, *T* being the time at which his actual decision occurs. It is also a necessary condition for Jones's deciding otherwise at any time after his deliberation process has begun. By "deciding otherwise," I mean any decision that is incompatible with his decision to break the promise; for example, the decision to stop deliberating.

Brain-Malfunction-*W* satisfies conditions (a) – (c) which define an IRR-situation. In it, the only alternatives to $D(-K)$ are alternatives which, though causally possible, are *actionally inaccessible* to Jones.

We are now in a position to answer the libertarian worry stated earlier (p. 166) regarding the alleged possibility of IRR-situations. The worry was: Why, given that an agent's decision is not causally determined, does he not have the power to avoid making it? The answer is this. The fact that the agent's decision is not causally determined entails that there is a causally possible alternative to it. But, as the above examples show, such an alternative need not be *actionally accessible* to the agent, that is, it need not be a state of affairs that the agent has the power to bring about.[14]

III. IRR-SITUATIONS AND FRANKFURT-TYPE EXAMPLES

As stated in the introduction, a Frankfurt-type counterexample to PAP purports to describe a situation in which: (i) an agent *S* decides-to-*V* on his own, but (ii) unknown to *S*, there is some factor that would have caused him to decide to *V*, had *S* not decided-to-*V* on his own. That is, a Frankfurt-type example purports to describe a situation in which there are just two options:

(2) Either *S* decides-to-*V* at *T* on his own, or *S* is caused by some factor to decide-to-*V* at *T*.

This feature is also shared by some of the examples in the recent literature that were aimed at improving upon Frankfurt's original example.[15] Now, (2) *looks* like an instance of IRR*. Hence, if coherent, a Frankfurt-type example would be a special case of an IRR-situation. Much of the recent debate between Frankfurtians and PAP-defenders has indeed focused on the question of whether the former have succeeded in describing a coherent scenario in which (2) is true. Since I have come to believe that none of the current attempts by Frankfurtians to describe such a scenario is successful, I shall (with one exception) not discuss them here again.[16] The exception is Pereboom's recent attempt to provide a Frankfurt-type example that works, as his example is considered by many to be a suitable example of an IRR-situation. After examining Pereboom's example, I shall return to the question of whether an adequate Frankfurt-type example can be constructed.

---

[14] One often hears the complaint that examples such as Brain-Malfunction-*W* or *Z*-type persons are of a *recherche* sort, and hence are uninteresting. I disagree. For all we know, neuroscientists may discover in the future that persons like those described in these scenarios actually exist.

[15] See some of the articles cited in footnote 4.

[16] My reasons for thinking so are based on responses to those attempts as mentioned in footnote 5.

Pereboom's example, which he calls "Tax Evasion 2" reads as follows:

> *Tax Evasion* (2): Joe is considering whether to claim a tax deduction for the substantial local registration fee that he paid when he bought a house. He knows that claiming the deduction is illegal, that he probably will not be caught, and that if he is, he can convincingly plead ignorance. Suppose he has a very powerful but not always overriding desire to advance his self-interest regardless of the cost to others, and no matter whether advancing his self-interest involves illegal activity. Crucially, his psychology is such that the only way that in this situation he could *fail* to choose to evade taxes is for moral reasons. His psychology is not, for example, such that he could fail to choose to evade taxes for no reason or simply on a whim. In addition, *it is causally necessary for his failing to choose to evade taxes in this situation that he attain a certain level of attentiveness to these moral reasons* [my emphasis]. He can secure this level of attentiveness voluntarily. However, his attaining this level of attentiveness is not causally sufficient for his failing to choose to evade taxes. If he were to attain this level of attentiveness, Joe could, with his libertarian free will, either choose to evade taxes or refrain from so choosing (without the intervener's device in place). More generally, Joe is a libertarian free agent. But to ensure that he choose to evade taxes, a neuroscientist now implants a device, which, were it to sense the requisite level of attentiveness, would electronically stimulate his brain so that he would choose to evade taxes. In actual fact, he does not attain this level of attentiveness, and he chooses to evade taxes while the device remains idle. In this situation, Joe could be morally responsible for choosing to evade taxes despite the fact that he could not have chosen otherwise.[17]

I see two problems with Pereboom's example. The first and main problem is that in it, the agent's decision is causally determined. To see this, let us examine the situation he describes *just before* (say one microsecond) Pereboom's agent (Joe) makes the decision to evade taxes. At that time, a causally necessary condition for his *not* making that decision (that is, his attaining the requisite level of attentiveness to moral reasons) is missing, which means that there is a causally sufficient condition for his making that decision; and at that time, it is no longer within Joe's power to change that. Voluntarily raising one's level of attentiveness to moral reasons is an intentional action that takes time, and at the point in time in question it would be too late for Pereboom's agent to complete it on time, that is, before the occurrence of his (actual) decision to evade taxes. Since the agent's decision is causally determined, the situation described is not an IRR-situation.[18]

---

[17] Pereboom, "Source Incompatibilism and Alternative Possibilities," in Widerker and McKenna, eds., pp. 193–95. For a similar example, see Hunt, "Moral Responsibility and Buffered Alternatives." I would respond to Hunt's attempt in much the same way as I am responding in what follows to Pereboom's attempt.

Another problem with Pereboom's example is that, in it, the agent is *derivatively* blameworthy for the decision he made, because he has not done his reasonable best (or has not made a reasonable effort) to avoid making it. He should have been more attentive to moral reasons than he in fact was—something he could have done. And in that case, he would not be blameworthy for deciding to evade taxes, as then he would be forced by the neuroscientist to so decide. If this is correct, then Pereboom's example is a case of derivative culpability, and hence is irrelevant to PAP, which, as stated earlier, concerns itself only with *direct* or *nonderivative* culpability (note 1). Pereboom seems to be aware of this criticism. He rejects it since he thinks that raising his attentiveness to moral reasons is not a morally significant (robust enough) option for Joe. He says:

> This example does feature alternative possibilities that are available to the agent—his achieving higher levels of attentiveness to moral reasons. But these alternative possibilities are not robust. First, note that in ordinary circumstances, without the intervener's device in place, it is not the case that by achieving some higher level of attentiveness Joe would have avoided responsibility for choosing to evade taxes. For under these conditions achieving some higher level of attentiveness is compatible with his not refraining from making this decision, or even ever being seriously inclined so to refrain, and choosing to evade taxes instead.[19]

I find this response unconvincing. As a morally competent agent who was aware of the temptation to evade taxes, Joe should have known that by raising the level of his attentiveness to moral reasons, he would be in a better position to overcome that temptation; and therefore, he should have availed himself of this option (something he could have done). To make this point clearer, suppose we consider a normal situation—one in which there is no neuroscientist in place—and in which Joe is caught by the IRS. At his trial, the judge reprimands him saying: "Joe, you should have reflected more on the immoral character of the deed you were about to perform, and should have refrained from doing it." Could Joe respond saying: "Sorry, Your Honor, but I am a person who upon becoming more morally aware as to what's at stake, would not even be inclined to refrain from evading taxes." Had Pereboom's agent answered the judge this way, the latter would be justified in inferring that he is dealing with an individual who has a character flaw. Hence, Pereboom's example is either a case of derivative culpability, or in case the agent is not culpable for the said flaw, it is an example in which the agent is not culpable for his decision.

---

[18] For other, somewhat more complex, criticisms of Pereboom's example, see Goetz, "Alternative Frankfurt-Style Counterexamples to the Principle of Alternative Possibilities," pp. 139–41, and Ginet, "Review of *Living without Free Will*," *Journal of Ethics*, VI (2002): 307–08.

[19] Pereboom, "Source Incompatibilism and Alternative Possibilities," pp. 194–95.

Is it possible to provide a successful Frankfurt-type example? Or alternatively put, is it possible to provide an example of an IRR-situation that takes the form of Frankfurt-type example? I am very skeptical that this can be done. The difficulty, as I see it, is this: a constraint on any Frankfurt-type example is that it allows for only two possible outcomes:

(2) Either *S* decides-to-*V* at *T* on his own, or *S* is caused by some factor to decide-to-*V* at *T*.

If so, then to guarantee that this constraint is satisfied, there must exist some prior condition *C* (indexed to a time *T** earlier than *T*) that is causally sufficient for (2).[20] But since (2) entails that

(3) *S* decides-to-*V* at *T*,

it follows that (3) must obtain in every causally possible world relative to *T**. However, if that is the case, it is hard to understand how (3) can fail to be causally determined; which means that we have every reason to believe that the scenario such an example describes is *not* an IRR-scenario. The same difficulty does not arise for examples of an IRR-situation we suggested earlier, since those are not subject to so stringent a constraint as the one stated in (2).

IV. A MORE GENERAL NOTION OF AN IRR-SITUATION

It is important to note that Frankfurt's case against PAP would be just as strong if instead of establishing the possibility of a situation in which Jones has no *actionally accessible* alternative to his decision not to keep his promise (IRR-situation), he would be able show that there might be a situation in which Jones has no *morally significant alternative* to that decision—for example, when Jones's only way of being able to avoid $D(-K)$ is by performing some act V^* that unknown to him would cause him to be in a state incompatible with $D(-K)$. For example, drinking a cup of water that is near him and that happens to contain a sleeping potion.[21] In such a situation, it would be counterintuitive to ground Jones's culpability for $D(-K)$ in the fact that he did not avail himself of the alternative possibility of drinking the said cup of water.[22] To capture this point better, we need a more general notion of an IRR-situation; one in which the agent does not necessarily lack an *actionally accessible* alternative to the morally wrong act he performs, but lacks rather a *morally significant alternative.*

---

[20] In the absence of such a prior condition *C*, one might claim that in a typical Frankfurt-example the agent's decision was avoidable, as then the two options mentioned in (2) would not be the *only* possibilities. There would also be the possibility of *S*'s refraining from the deciding-to-*V* at *T*.

Call an act V^* a *morally significant alternative* to a morally wrong act V that an agent S performs at T if and only if: (i) V and V^* are incompatible, (ii) S is aware (or should be aware) that if he performs V^* at T, he will not be acting in a morally wrong way, (iii) S truly believes that it is within his power to perform V^* at T, and (iv) in the circumstances, V^*-ing at T would be regarded as a reasonable way for S to avoid V-ing at T.[23]

We can now introduce the following more *general* notion of an IRR-situation:

> (G-IRR) There may be circumstances that in no way bring it about that a person performs a certain action; nevertheless those very circumstances leave him with no morally significant alternative to performing that action.[24]

---

[21] Cf. Pereboom, *Living without Free Will*, pp. 25–26.

[22] John Fischer was the first to draw attention to such alternatives and their bearing on PAP. He refers to them as "alternatives that are not sufficiently robust for ascriptions of responsibility"—see *The Metaphysics of Free Will* (Cambridge: Blackwell, 1994), chapter 7.

[23] The need for the first three clauses is obvious. The fourth clause is needed to exclude from the range of morally significant alternatives acts which, though satisfying (i) – (iii), would not normally be regarded as reasonable ways for S to avoid doing V (for example, deliberately committing suicide, or deliberately taking a sleeping pill, and so on). If doing V is deciding to break the promise to Smith, and Jones was justified in believing that (in the circumstances) he could have avoided doing so simply by deciding *not* to break it, then there was no need for him to go so far as to take a sleeping drug in order to avoid deciding to break the promise. Just a reasonable effort on his part was required. For a closely related, but still incomplete definition of a morally significant alternative, see Pereboom, "Source Incompatibilism and Alternative Possibilities," p. 188. For a related version of this principle which I also do not accept, see McKenna, "Robustness, Control, and the Demand for Morally Significant Alternatives," p. 209.

[24] Corresponding to this more general formulation of an IRR-situation would be the following more general formulation of PAP:

> An agent S is morally blameworthy for V-ing at T only if S had a *morally significant alternative* to V-ing at T.

As the notion of a G-IRR-situation is more general than that of an IRR-situation, the task of constructing an example of a G-IRR-situation should be even easier than that of constructing an example of an IRR-situation.[25] In what follows, I shall, for simplicity's sake, ignore the distinction between an IRR-situation and a G-IRR-situation, and continue to speak only about the former.

V. DIALECTICAL INTERLUDE

Suppose that IRR-situations are indeed possible. What is the significance of this result? First, there is an important metaphysical implication. Such scenarios force libertarians to distinguish clearly between two basic notions or senses of acting freely—the "on my own" sense and the "up to me" sense. According to the first sense, acting freely consists in performing an action[26] without being nomically determined or caused to perform it, whereas according to the second sense, acting freely consists in an agent's performing an action such that it was within the agent's power to avoid performing it. This is in sharp contradistinction to the traditional libertarian position, which, though acknowledging the conceptual difference between the two senses, regarded them as (broadly) logically equivalent.[27]

Next, what about the *moral* significance of IRR-scenarios, or more specifically, their bearing on an adequate account of moral blame? Do they imply that PAP is false? In the past, I have resisted this conclusion pointing out that to establish it we also must assume that

(M) In an IRR-situation, the agent is blameworthy for the morally wrong act he performed.

This assumption, I claimed, is far from being obvious.[28] To show this, I drew attention to a powerful consideration against (M)—the intuitive link between attributing blame to an agent and expecting of the agent that he not have done what he did. When we consider someone morally blameworthy for a certain act, we believe that morally speaking he should *not* have done what he did, and we blame him because he has not met that expectation.[29] Sometimes, however, such an expectation would be unreasonable. For example, in a situation in which it is clear to us that the agent could not have avoided acting as he did. To expect of the agent in that situation not to have done what he did would be to demand of him the impossible, and would, therefore, be irrational. This consideration led me to suggest the following necessary condition for moral blame which I called the "Principle of Alternative Expectations":

---

[25] For a worked out example, see Widerker, "Blameworthiness, Non-Robust Alternatives, and the Principle of Alternative Expectations," section 1.

[26] By "action" I understand an action in Ginet's sense, that is, either a causally simple mental action such as a volition-to-*V* or the forming of an intention-to-*V*, or a complex action such as Sam's action of killing Smith that consists of a simple mental action causing the event of Smith's death. Crucial here is the idea that what makes a certain event an action is something that is intrinsic to it, and not the way the event is caused. See Ginet, *On Action* (New York: Cambridge, 1990), chapter 1.

[27] Among those who draw the same moral from the conceptual possibility of IRR-situations are: Linda Zagzebski, "Does Libertarian Freedom Require Alternative Possibilities?" *Philosophical Perspectives*, XIV (2000): 231–48; Stump, "Alternative Possibilities and Moral Responsibility: The Flicker of Freedom"; Hunt, "Moral Responsibility and Unavoidable Action"; and Pereboom, *Living without Free Will*, chapter 1.

[28] See Widerker, "Blameworthiness and Frankfurt's Argument against the Principle of Alternative Possibilities," in Widerker and McKenna, eds., pp. 63–67.

> (PAE) An agent *S* is morally blameworthy for performing a given act *V* only if in the circumstances, it would be morally reasonable to expect of *S* not to have done it.[30]

This principle, I argued, explains why one may want to reject (M). Moreover, it provides the rationale for a principle such as PAP, which, in the light of PAE, may be viewed as a reasonableness constraint on ascriptions of moral blame. Here, I do not intend to strengthen my case for PAE further. Instead, I wish to draw attention to a libertarian-based alternative to PAE, that differs importantly from Frankfurt's account of blame, but which like the latter does not regard avoidability as a necessary condition for moral blame. I shall dub such an account an "avoidability-independent account of moral blame" or AI-BL. My goal is to show that, just as there is something very appealing about PAE, there is also something very intuitive about the claim (M) that in an IRR-situation the agent is blameworthy for what he did, and that this intuition can be captured in libertarian terms. I shall, however, not adjudicate between this alternative account of blame and PAE, but leave it for the reader to decide which is preferable.

VI. A LIBERTARIAN-BASED ALTERNATIVE TO PAE

Before developing the said alternative to PAE, let us first ask ourselves why we tend to regard an agent who acts wrongly in an IRR-situation as blameworthy, even though that he could not have avoided acting as he did? Surprisingly, this question has only received scant attention in recent discussions of Frankfurt's argument against PAP. Most discussants, with the notable exception of Frankfurt himself, either just content themselves with the observation that it is intuitive to think so, or they adopt Frankfurt's answer to that question stated below:[31]

---

[29] Cf. R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge: Harvard, 1994), chapter 4.

[30] Like PAP, PAE is not meant apply to cases of derivative culpability; see note 1. By "morally reasonable" in PAE, I mean morally reasonable for someone who is morally competent and knows all the relevant nonmoral facts pertaining to the situation the agent is in.

> The fact that a person could not have avoided doing something is a sufficient condition for his having done it. But as some of my examples show, this fact may play no role whatever in the explanation of why he did it. It may not figure at all among the circumstances that actually brought it about that he did what he did.... Now if someone had no alternative to performing a certain action but did not perform it because he was unable to do otherwise, then he would have performed exactly the same action even if he *could* have done otherwise. The circumstances that made it impossible for him to do otherwise could have been subtracted from the situation without affecting what happened or why it happened in any way.
>
> ....When a fact is in this way irrelevant to the problem of accounting for a person's action *it seems gratuitous to assign it any weight in the assessment of his moral responsibility* [my emphasis]. Why should the fact be considered in reaching a moral judgment concerning the person when it does not help in any way to understand either what made him act as he did or what, in other circumstances, he might have done? (*op. cit.*, pp. 836–37)

Put succinctly, Frankfurt's argument in support of the claim that

(M) In an IRR-situation, the agent is blameworthy for the morally wrong act he performs,

is the following:

(1) The fact that in an IRR-situation an agent could not have avoided performing a certain act plays no role in the causal explanation of why he performed it.
(2) If a fact is irrelevant to the causal explanation of why the agent performed a certain act, then such a fact has no bearing on the agent's moral responsibility for the act.
(3) Therefore, (M) is the case, and PAP is false.

The problem with this argument lies in its second assumption. There are intuitive examples that show that sometimes the reason why we absolve an agent from blame (or hold him blameworthy) for performing a certain act, does not figure at all in the causal explanation of the act. Here is one such example. Suppose that *S* wakes up one morning with an irresistible urge to strangle his neighbor's canary. Despite all his efforts to resist that desire, he ultimately succumbs to it. Unbeknownst to *S*, he has been manipulated to so act. Clearly *S* is not blameworthy for strangling the canary, and the reason for this is that he did all he could to refrain from doing so. But note that this fact is irrelevant to the explanation of why he performed that act. He killed the canary not because of his efforts to resist, but despite them.[32] Here is another counterexample to (2). Suppose that *S* harms another person for some selfish reason, knowing very well that in doing so he is acting immorally. Surely the fact that *S* knew that he was wrong to do what he did has a bearing on our considering him blameworthy for his act. But again this fact plays no role in the causal explanation of why *S* performed that act. He harmed the other person not *because* he knew that doing so is morally wrong, but despite his knowing that.[33]

---

[31] By "those discussants" I have in mind Stump, Hunt, Pereboom, Mele, and others.

I believe that there is a better answer to the question of why we tend to regard an agent who acts in a morally wrong way in an IRR-situation as blameworthy. To see it, we need to describe the sequence of events from the agent's point of view. Keeping this in mind, let us look again at some particular example of an IRR-scenario, say, Brain-Malfunction-*W* (see p. 169). Here Jones sees himself as facing two options: accepting Mary's offer, thereby breaking the promise he made to his uncle, or rejecting her offer and keeping the promise. After deliberating about what to do, he *chooses on his own* to realize the one that he believes to be morally wrong. I say "on his own," because (ex hypothesi), the decision he makes is his *own* in the libertarian sense in that it is not nomically necessitated or caused. I say "chooses" because he makes that decision in the belief that he could have decided otherwise.[34]

It is clear now why we find Jones's behavior in Brain-Malfunction-*W* morally objectionable. We do so because in deciding not to keep the promise—something Jones was aware is morally wrong—*he chose on his own* to violate his obligation to keep his promise in favor of something which from a moral view point counts as less important, namely, his selfish interest to spend time with Mary. This reason for his being blameworthy can be also expressed more elaborately as follows: Before deciding not to keep his promise, Jones was aware that doing so would be morally wrong. Hence, *he had a reason* not to act this way. Furthermore, he did not think (nor had he a good reason to think) that deciding to keep the promise was not within his power to do. Nor did he believe that acting in this way would be irrational.[35] This means that by deciding not to keep the promise he *chose on his own* to act contrary to the way required by moral reasons. Therefore, he is *morally to* blame for what he did.

---

[32] This example is a variation on an example by Bernard Berofsky in his *Freedom from Necessity* (New York: Routledge and Kegan Paul, 1987), p. 42.

[33] Recently, Frankfurt has conceded that (2) is false. He now insists that the fact that *S* was aware that he was acting immorally is part of what he calls "a morally pertinent explanation of his act" by which he means an explanation of why *S* acted immorally. See his "Some Thoughts Concerning PAP," in Widerker and McKenna, eds., pp. 339–45. But if this is what Frankfurt means by "explanation," then assumption 1 of his argument for (M) begs the question against the proponent of PAP.

[34] This is consistent with the fact Jones could not have chosen otherwise, since choosing only requires that one acts *in the belief* that one could have acted otherwise.

Note that the fact that Jones mistakenly believed that he could have decided otherwise, but in fact could not have done so, is morally irrelevant. It does not change a bit our negative assessment of him as someone who chose on his own to act contrary to moral reasons. Jones is just as morally unjustified for acting as he did, as he would have been had he had the option of deciding otherwise. It is this point that I take to be one of the important lessons to be drawn from the possibility of IRR-scenarios.

The above considerations suggest the following account of moral blame:

(AI-BL) An agent S is morally blameworthy for deciding-to-V at time T if and only if

(i) S decides-to-V at T on his own.
(ii) S decides-to-V at T being aware that doing so is morally wrong.
(iii) S believes that he could have decided otherwise, or believes that he has no good reason to think that he could not have decided otherwise.
(iv) S does not think that, in the circumstances, deciding-not-to-V at T is irrational.[36]

---

[35] He would be justified in believing so if, prior to his decision to stay with Mary, he were to be informed that due to a terrorist attack all the roads to the hospital where his uncle is are blocked.

[36] What if S is a religious fatalist? Does it follow (as AI-BL seems to imply) that *just* because he believed that he could *not* have decided otherwise he is to be taken off-the-hook? That depends on whether that belief was deliberatively efficacious, that is, played a role in the deliberation process leading S to decide as he did. If it did play such a role, then (unless S is culpable for holding that belief) S should be exonerated. Alternatively, if S's belief in fatalism was deliberatively inert, then S is culpable for the decision he makes. To do justice to this point clause (iii) of AI-BL would need to be modified as follows: (iiia) Either S believes that he could have decided otherwise, or believes that he has no good reason to think that he cannot do otherwise, or if S believes that he cannot decide otherwise, then (in the circumstances) that belief was deliberatively inefficacious with regard to S's decision to V. My thanks to Michael Feld for drawing my attention to the case of a religious fatalist.

Now someone who satisfies the conditions stated above cannot in order to exonerate himself from blame appeal to regular excuses such as ignorance, coercion, hypnosis, and the like. Nor can he view himself as morally justified for the way he acted. And so we can state the above account more generally as follows:

> An agent *S* is blameworthy for *V*-ing at *T*, if and only if
>
> (i) *S V*-s at *T* on his own, and
> (ii) *S* is aware that by *V*-ing at *T* he violates a (prima facie) moral obligation without having a morally proper justification or excuse for so doing.

We may refer to such a person as "someone who, by acting as he does, shows a lack of respect (concern) for morality."

The account before us is an avoidability-independent account of moral blame.[37] According to it, the agent's blameworthiness does not derive from the fact that he did *not* avail himself of an alternative way of acting that he *should have* and *could have* availed himself of. Rather, the agent is blameworthy because by acting as he did, he expressed a lack of respect (concern) for morality in the sense specified above. Note, however, that *pace* Frankfurt, the "could have done otherwise" notion of acting freely still figures in it importantly, not as a metaphysically necessary condition, but as an epistemic one.[38]

VII. THE ABOVE ACCOUNT OF BLAME AND FREE WILL

Before turning to considering some potential objections to the above account of moral blame, let us reflect once more on how free will figures in it. It does so in two ways involving two different senses of "free act." First, the agent's act must be free in the sense that he did it *on his own*. This is so, since for libertarians an agent cannot be *directly* blameworthy for doing something if he did not do it on his own. Second, the agent's act must be *epistemically* free (epistemically up to him), in the sense that the agent performed it believing that he could have acted otherwise (or believing that he had no good reason to think that he could not have acted otherwise). Only if this second condition is satisfied can the agent be said to have *chosen* to act contrary to the way required by moral considerations, thus indicating that by so acting he was showing a lack of respect (concern) for morality.[39] Together these two conditions form the notion of *choosing on one's own* to act in a certain way, which for the proponent of AI-BL constitutes the freedom-relevant condition for moral blame. Call them the "metaphysical" and the "epistemic component" of that notion. Note that the two components are logically independent. The fact that an agent acted *on his own* does not preclude the possibility of his falsely believing that he was in an IRR-situation, or his falsely believing that his act was caused by Allah. Alternatively, an agent may be manipulated into thinking that he is acting on his own, even though he himself believes that he could have acted otherwise.

---

[37] Fischer calls such an account an "actual sequence account" of moral blame. See Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge, 1998), chapter 2; and Fischer, "Recent Work on Moral Responsibility," *Ethics*, cx (1999): 93–139, p. 125.

[38] Note that this account of blameworthiness would be also congenial to a compatibilist, though of course, the latter would construe such notions as "acting on one's own," and "could have done otherwise," in a compatibilist way.

Now "choosing on one's own to act in a certain way" in the sense explained above is a distinct and important sense of acting freely. The significance of IRR-scenarios from a libertarian viewpoint is that they draw our attention to this specific sense, and to the fact that this sense cannot be accounted for in terms of avoidability.

VIII. OBJECTIONS AND REPLIES

I would like to conclude this paper by considering some objections that may naturally arise in connection with the above account of blame.

Objection 1: Holding Jones blameworthy for deciding not to keep his promise in a situation such as Brain-Malfunction-*W*, when, in fact, he could not have avoided doing so (or could not have avoided being blameworthy), is unfair to Jones.

Reply 1: The objection is unconvincing. Why exactly is this unfair? If the reason is that: Jones *could not* have done what he *should* have done but did *not*, then the critic assumes that if Jones is blameworthy, then the reason for this being so, is that he did *not* do what he *should* and *could* have done. Note, however, that this is not how the proponent of the AI-BL account of blame is viewing the situation. He does not take *that* to be the reason why Jones is blameworthy. For him, Jones is blameworthy because the decision Jones made in Brain-Malfunction-*W* expressed a lack of respect for morality on his part.

What about other reasons for thinking that it is unfair to blame Jones for deciding not to keep the promise, when he could not have avoided doing so? Can one say that, if Jones had the option of deciding to keep the promise, he would be morally better off?

---

[39] If the agent did not believe that he could have acted otherwise, or had a good reason for thinking that he could not have acted otherwise, then he would have had an excuse, and would not be blameworthy for what he did.

The answer is *no.* For we have no good reason to think that he would have acted differently if he had that option. But perhaps the reason why that is unfair is because then we would be treating Jones and his non-Frankfurtian counterpart (who could have avoided being blameworthy) equally, even though the latter had an option Jones did not have? Again, the answer is negative. The reason for holding Jones blameworthy is that he acted in a morally wrong way in the absence of a morally justified reason for doing so,[40] and in that regard he does not differ from his non-Frankfurtian counterpart.

Objection 2: If we accept the AI-BL account of blame we have to give up the conceptually intuitive link between holding an agent blameworthy and expecting of him not to have acted as he did.

Reply 2: True. But note that if we adopt the "reasonableness of expectation" account of moral blame (PAE), we also have to give up certain basic intuitions about moral blameworthiness. For example, we have to give up the intuitive assumption that: if *S* has a moral obligation to *V* and violates it in the absence of having an adequate justification or excuse, then he is blameworthy for so acting. This assumption is often viewed as being explicative of the very notion of moral obligation.[41] So just as there are costs to abandoning an account of moral blame such as PAE, there are parallel costs to insisting that in an IRR-situation the agent is not blameworthy. Ultimately, deciding which account of blame is to be preferred will have to involve a whole array of considerations, and it is not clear at all that at the end of the day the traditional non-Frankfurtian account of moral blame will turn out to be the better one.[42]

Objection 3: Holding Jones blameworthy for deciding not to keep his promise in a situation like Brain-Malfunction-*W* is irrational and hence, morally inappropriate. For what should he have done instead, or what should he have done so as to avoid being blameworthy? One cannot say that he should have avoided deciding to break the promise, since it was not within his power to do that.

---

[40] By "morally justified reason" I mean "justification" or "excuse" in J.L. Austin's sense. See Austin, *Philosophical Papers* (New York: Oxford, 1970), pp. 157–76.

[41] See, for example, Kurt Baier, "Moral Obligation," *American Philosophical Quarterly,* III (1966): 210–26; Marina Oshana, "Ascriptions of Responsibility," *American Philosophical Quarterly,* XXXIV (1997): 76–77; and Stephen Darwall, "Moral Obligation and Accountability," forthcoming.

[42] For other advantages of an avoidability-independent account of moral blame over PAE, see Ken Levy, "Why It Is Sometimes Fair to Blame Agents for Unavoidable Actions," *American Philosophical Quarterly,* XLVII (2005): 93–104.

Reply: This is what I have referred to elsewhere as "the W-defense,"[43] and it strikes me as one of the strongest challenges to the account of blame under consideration. Still, a defender of that account can reply to it in the following way: In insisting on pressing the question: "What should Jones have done instead, or what should he have done so as to avoid being blameworthy?" the critic presupposes that in order to be blameworthy one must have had the option of not being blameworthy. But this assumption is rejected by the proponent of the AI-BL account of blame. On that account, an agent is blameworthy if and only if by acting as he did he showed a lack of respect (concern) for morality. To act in this way, it is not necessary that the agent had (in fact) the option of not being blameworthy. It is enough that he *believed* that he had it. And this is certainly true of Jones who, when making the decision to stay with Mary, believed that he could have decided otherwise.

Perhaps the following remark will make the above reply clearer. Consider again the question "What should Jones have done instead, or what should he have done so as to avoid being blameworthy?" This question, the proponent of AI-BL may explain, is similar in an important respect to the question: "When did you stop beating your wife?" when posed to someone *S* who is a good husband. The fact that *S* does not have a good (direct) answer to it, certainly does not entail that *S* is to blame for having been cruel to his wife in the past. The proper response to it by *S* is an indirect one. It is to draw the questioner's attention to the fact that in asking the question he is presupposing that *S* used to beat his wife in the past, and then to point out that this presupposition is not true. In a similar way, (so argues the proponent of AI-BL) the question: "What should Jones have done instead, or what should he have done so as to avoid being blameworthy?" also carries with it a presupposition, namely, that to be blameworthy, one must have had the option of not being blameworthy. And this is precisely what the proponent of the account of blame under consideration is denying.

Objection 4: Here is an objection, closely related, to the previous one. If Jones is blameworthy for deciding to break the promise, then this implies that he *should not* have decided to act in this way. But how can this be the case if it was not within his power to refrain from deciding to break the promise? On the other hand, if Jones did *not* have the obligation to refrain from deciding to break the promise, it follows that the proposed account of blame has the odd consequence that an agent can be blameworthy for *V*-ing, even if he did not have an obligation to refrain from *V*-ing.

---

[43] See Widerker, "Frankfurt's Attack on the Principle of Alternative Possibilities: A Further Look," p. 191, and "Blameworthiness, Nonrobust Alternatives, and the Principle of Alternative Expectations," pp. 296–98.

Reply: The first part of the above objection is based on the assumption that "ought implies can," an assumption regarded by many as problematic.[44] But even if one grants that assumption and agrees that Jones did not have the obligation to refrain from breaking the promise, the objection still does not go through. Note that if Jones's blameworthiness derives from the fact that by acting as he did he expressed a lack of concern for morality, then to determine whether he is blameworthy, we have to assess his conduct not relative to what his obligation *in fact* was, but relative to what he took it to be. And here the facts are quite clear. Jones believed that he was morally obligated to decide to keep the promise, and that he did not have a morally justified reason not to do so. Nonetheless, he chose to act contrary to what he believed morality required. In so acting he has shown that he cares more for his selfish interests than for those of morality. And for this he deserves to be regarded blameworthy. All this, of course, modulo the assumption that what he did, he did on his own.

Objection 5: The account of blame under consideration does not enable us to distinguish someone like Jones who is blameworthy for what he did, from his morally conscientious counterpart Jones* who, were he able to decide otherwise, would have refrained from breaking the promise he made. For, if in Brain-Malfunction-*W* the option of deciding otherwise is closed off, then Jones* may also end up deciding not to keep his promise.

Reply: It is true that in a scenario such as Brain-Malfunction-*W* Jones*, too, may end up deciding not to keep his promise. Still, there is an important difference between him and Jones. Unlike Jones, Jones* would have decided differently in a normal situation in which the neurologically necessary condition for his deciding otherwise, namely, *N*, were in place. This implies that, if in a scenario such as Brain-Malfunction-*W*, Jones* were to decide to break the promise, he

---

[44] See for example, Peter Graham, "'Ought' Does *Not* Imply 'Can'," Unpublished Manuscript; Walter Sinnott-Armstrong, *Moral Dilemmas* (New York: Blackwell, 1988), p. 113; Alan White, *Modal Thinking* (Ithaca: Cornell, 1975), p. 152; and E.J. Lemmon, "Deontic Logic and the Logic of Imperatives," *Logique et Analyse*, VIII (1965): 47–50. Note that the critic cannot take "he *should not* have decided to act in this way" to mean that it was reasonable to *expect* of the agent *not* have decided to act in this way. For then he would be begging the question against the proponent of AI-BL.

(unlike Jones) would do so *because of* the absence of *N*. But then the pressure for holding Jones* blameworthy vanishes, since for him Brain-Malfunction-*W* would not be an IRR-situation.[45]

CONCLUSION

The avoidability-independent account of moral blame to which I have been calling attention here places me in the camp of those called "Source Incompatibilists" who hold that, although moral responsibility is incompatible with determinism, moral responsibility does not require that the agent could have avoided acting as he did.[46] On this view, moral responsibility just requires that the agent was the ultimate originator of his act, that is, he performed the act without being in any way caused or nomically determined to perform it. Perhaps a better term for such a position would be "Frankfurt-friendly libertarianism."[47] The difference between my position and that of other Source Incompatibilists consists in the kind of examples I use to show that IRR-scenarios are conceptually possible, and the reasons I give as to why such scenarios justify the rejection of PAP.

One final remark as to how my position on IRR-scenarios and their moral significance relates to that held by Frankfurt. Like Frankfurt, I believe that IRR-situations are conceptually possible. However, I disagree with his reasons for this contention, or more specifically with the kind of example that he uses to establish it. As for the account of blameworthiness suggested by those scenarios: First, as I explain in section v, I do not wish to adjudicate between that account and the "reasonableness of expectation" account I have defended elsewhere. I still find the intuition behind the latter account powerful and deeply entrenched in our thinking about moral blame. That said, I do acknowledge the intuitive force of Frankfurt's claim that the agent in an IRR-situation is blameworthy for what he did, even though that he could not have done otherwise. But here, too, my way of accounting for this intuition differs from Frankfurt's. Frankfurt sees the reason for the falsity of PAP in the fact that the unavoidability of the agent's act (the factor that renders it unavoidable) did not play any role in the causal explanation of the act. I, on the other, hand, attribute the failure of PAP to the fact that the agent, in acting as he did in an IRR-situation, showed a lack of respect for morality (that is, chose on his own to act contrary to the requirements of morality without being morally justified in doing so), and to the observation that this negative verdict would still apply even if one were to assume that the agent's act was avoidable. However, the difficulties I find with Frankfurt's position in no way detract from the importance of Frankfurt's achievement, that is, that by drawing our attention to the possibility of IRR-scenarios and their moral implications, Frankfurt has profoundly deepened our understanding of the notions of freedom and responsibility.

---

[45] Here is a problem that I leave to those sympathetic to an AI-BL account of blame to consider: There are certain attitudes/sentiments that are closely associated with the notion of blame. For example, we believe that

> If *S* is aware that he acted culpably, it is appropriate for *S* to express remorse for his behavior.

Note, however, that according to the AI-BL account of blame this rather intuitive assumption would be false. To see this, suppose that, after having broken the promise to his uncle, Jones were to discover that the situation he was in was an IRR-situation, that is, that he could not have avoided deciding to break the promise. Reflecting upon his deed and being inclined to express remorse for it, Jones would immediately realize that it would be irrational for him to do so. For the belief that he should not have acted as he did—a belief essential to the expression of remorse—would be defeated by his other belief that in the circumstances he lacked an actionally accessible alternative. The proponent of the AI-BL-account is thus left with the problem of having to explain why, if Jones acted culpably, it is inappropriate for him to express remorse. This objection attests to how deeply entrenched the PAP-intuition is in our conception of blame.

[46] For example, Zagzebski, "Does Libertarian Freedom Require Alternative Possibilities?"; Stump, "Alternative Possibilities and Moral Responsibility: The Flicker of Freedom"; Hunt, "Moral Responsibility and Unavoidable Action"; and Pereboom, *Living without Free Will*, chapter 1.

[47] For this handy term, I am indebted to Ira Schnall.

DAVID WIDERKER

Bar-Ilan University

CONTROL, CHOICE, AND THE CONVERGENCE/DIVERGENCE DYNAMICS: A COMPATIBILISTIC PROBABILISTIC THEORY OF FREE WILL*

A theory of free will needs to account for the type of control that human agents exercise in free actions. Understanding this type of control is important for central issues in moral philosophy, such as those concerning responsibility and desert; for one can hardly be thought to be morally responsible for actions one does not control.[1] Yet a proper understanding of control is missing in both compatibilist and libertarian theories. Is the existence of alternative possibilities for actions, necessary or sufficient for this type of control? If it is neither, would there be any way to explain in what way control over actions is consistent with causal determination but is diminished in situations that involve *covert nonconstraining* (CNC) compulsion, such as Skinnerian indoctrination or value engineering? As observed by Robert Kane (*op. cit.*), the clashing intuitions of some of the parties in this debate have brought the field to a somewhat deadlocked position.

Here I argue that the control needed for exercising the type of free will required for responsibility and autonomy does not depend on whether determinism or indeterminism is true. Instead, I will try to show that the critical element for the presence of responsibility and autonomy is a more complex dynamical property involving a succession of convergence and divergence dynamics. The convergence dynamics (attractors) will be shown to enable teleological (goal directed) behavior, while the divergent dynamics involve bifurcations in

---

* Special thanks to Nick Zangwill for comments on successive drafts of the paper and countless discussions and suggestions that helped me develop and clarify these ideas. I also want to thank Ariel Kernberg, Eytan Ruppin, Orly Shenkar, and Dan Tidhar for very stimulating discussions and comments, and Alfred Mele and Saul Smilansky for a critical reading of the manuscript. A simplified version of some of these ideas in the context of neurocomputational models was presented in a technical report (Marius Usher and Eytan Ruppin, "Free Will in Light of Chaotic Systems," TR 197/91 (1991), Institute of Computer Science, Tel-Aviv University, Israel).

[1] For good discussions, see Randolph Clarke, "Incompatibilist (Nondeterministic) Theories of Free Will" (The Stanford Encyclopedia of Philosophy, 2005); Robert Kane, *The Significance of Free Will* (New York: Oxford, 1996); John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control* (Cambridge: Blackwell, 1994); Alfred Mele, *Autonomous Agents: From Self-control to Autonomy* (New York: Oxford, 1995); Saul Smilansky, *Free Will and Illusion* (New York: Oxford, 2000).

0022-362X/06/0304/188–213

choice and in character formation. This opens the door to a compatibilist theory that does not clash with powerful intuitions of open-ended choice in deliberation.[2] In my approach I use a probabilistic framework for two reasons. First, *physical* laws may be indeterministic, in which case the causal relation between the brain states involved in choice and the ensuing actions is probabilistic. Second, even if the *physical* laws are deterministic, it does not follow that mental states are sufficient for uniquely determining decisions and actions, due to the way that the mental states are realized in physical states and to the probabilistic nature of the environmental causes. This will involve appealing to the notion of *constitutive luck*,[3] which is central to this account.

I start by examining some aspects of free will that an agent needs to possess in order to be responsible for her actions, focusing on the question of whether she needs to be able to act otherwise, and on some restrictions posed by psychological overdetermination (impatient readers who are familiar with these issues can jump to the next section). I then deploy a theory based on a mechanism of *guidance control* that enables teleological behavior, and which assumes that action selection and control involve a succession of bifurcation/attractor states. I argue that the theory solves a number of puzzles about autonomy and CNC-compulsion.

I. THE PRINCIPLE OF ALTERNATIVE POSSIBILITIES

A central issue in the current debate between compatibilists and libertarians is whether there must be *alternative possibilities* for actions that we are morally responsible for. Let us call the statement that alternative possibilities are necessary for responsibility the *principle of alternative possibility* (PAP). A key issue in the debate on PAP is Harry Frankfurt's counterfactual scenario, where, although there is no alternative possibility to an action, once the action is performed we have a compelling intuition that the agent is responsible.[4] The scenario involves a counterfactual intervener, who overrides the agent's libertarian choice (say, by directly activating her brain choice centers), only if she showed a *sign* of choosing *B* but not, in the actual case, when she chooses *A*. In the actual case, as no external intervention took place and the agent carried her choice and action on her own, we feel that the agent is responsible despite the fact that the alternative action (or even choice) was blocked by the counterfactual intervener. This argument is open to the objection that despite the intervener's effort, the agent still maintains some *flicker of freedom*; she may still *try* to choose otherwise and much debate has focused on the question of whether such a flicker of freedom is robust enough to ground responsibility.[5] For example, if the sign (or its absence) is deterministically related to the available options, *A* or *B* (as is needed if the contravener is to prevent the alternative action on the basis of the sign), one may argue that the presence (or absence) of the sign *is*, in effect, the choice, and thus the counterfactual intervention comes too late; the agent *does* have a robust alternative after all. More recently, a more refined counterfactual intervention scenario has been presented by Derk Pereboom.[6]

---

[2] For eloquent illustrations of the difficulty to give up on these intuitions, see Thomas Nagel, *The View From Nowhere* (New York: Oxford, 1986); John Searle, "Free Will as a Problem in Neurobiology" (talk given at Royal Institute of Philosophy in February 2001, based on "Consciousness, Free Action and the Brain," *Journal of Consciousness Studies*, VII, 10 (October 2000): 3–22).

[3] Nicholas Rescher, *Luck: The Brilliant Randomness of Everyday Life* (Pittsburgh: University Press, 1995).

[4] Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 1969): 829–39.

Although some libertarians insist on the flicker of freedom objection, I believe that robustness considerations in such scenarios give us a good reason to prefer an account of responsibility that does not rely on PAP, unless an independent argument in favor of PAP convinces us otherwise. One of the best articulated arguments in support of PAP is the *no-matter-what* principle advanced by Peter van Inwagen.[7] According to it, "if it is a fact that *p*, an agent is morally responsible for the fact that *p* only if that agent was once able to act in such a way that it would not have been the case that *p*" (*ibid.*, p. 376). To support the *no-matter-what* principle, van Inwagen presents a number of variants of Frankfurt scenarios, but without counterfactuals, such as:

---

[5] For detailed discussions, see Fischer, *The Metaphysics of Free Will*, and his "Recent Work on Moral Responsibility," *Ethics*, CX (October 1999): 93–139; Kane; Eleanor Stump, "Libertarian Freedom and the Principle of Alternative Possibilities," in Daniel Howard-Snyder and Jeff Jordan, eds., *Faith, Freedom and Rationality* (Totowa, NJ: Rowman and Littlefield, 1996), pp. 73–88; Stewart Goetz, "Stumping for Widerker," *Faith and Philosophy*, XVI (1999): 83–89; Mele, "Soft Libertarianism and Frankfurt-style Scenarios," *Philosophical Topics*, XXIV, 2 (1996): 123–41; Mele and David Robb, "Rescuing Frankfurt-style Cases," *Philosophical Review*, CVII (1998): 97–112; David Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *Philosophical Review*, CIV (1995): 247–61.

[6] Pereboom, "Alternative Possibilities and Causal Histories," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 119–37. In the scenario he imagines, the sign is a moral thought in favor of *B*, which occurs indeterministically and is *necessary* for the choice of *B*, but does not deterministically cause it; the agent may still indeterministically choose either *A* or *B*. If the agent chooses *A* (on her own), because the moral reason did not occur and thus no intervention took place, the agent is responsible, and the alternative possibility—in which the moral thought occurred—does not seem robust enough to ground the agent's responsibility. This is because, unlike in the original scenario, the sign does not settle the choice: despite the occurrence of the moral thought the agent may still decide against it.

[7] Van Inwagen, "Fischer on Moral Responsibility," *The Philosophical Quarterly*, XLVII, 188 (July 1997): 373–81. A similar principle, the *W*-principle (for "what-else-could-I-

> (S_1) "I am supposed to take the serum upriver to the plague-driven village. But I get drunk and miss the boat. Taking the boat is the only possible way to get to the village. Soon after the boat leaves the dock, it strikes a rock and sinks. Hundreds of villagers who would have been saved by the serum die" (*ibid.*, p. 378).

Van Inwagen argues that the agent cannot be held responsible for the death of the villagers (though he acted *irresponsibly* and is guilty of dereliction of duty) for the obvious reason that he could not have saved them anyhow. Furthermore, the same conclusion should apply, according to van Inwagen, to the counterfactual Frankfurt scenarios. To evaluate the *no-matter-what* principle, consider a situation involving causal overdetermination. The following scenario is modified (to include overdetermination) from van Inwagen.

> (S_2) Gunnar and Rifler desire Ridley's death and plan, independently, to murder him. It so happens that, without knowing of each other, both of them shoot Ridley at exactly the same moment. The bullet shot from Gunnar's gun and the bullet shot from Rifler's rifle hit Ridley in his chest and kill him. A pathological investigation shows that Ridley was still alive when both of the two bullets hit him (it took him few seconds to die) and that each of the bullets would undoubtedly and independently have caused his death. Are either Gunnar or Rifler "morally responsible for the fact that Ridley died and his children are now orphans" (*ibid.*, p. 379)?

The interesting feature of this scenario, is that by appealing to the no-matter-what principle, both Gunnar and Rifler can claim *not* to be responsible for Ridley's death. Whatever Gunnar had chosen, Ridley would still have died as the result of the other bullet. But this clearly leads to a highly counterintuitive conclusion that no one is respon-

---

do"), was advocated by Widerker, "Frankfurt's Attack on the Principle of Alternative Possibilities: A Further Look," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 181–201. I agree with Fischer's criticism that in the absence of independent evidence, this principle begs the question against the compatibilists, who insist that the relevant issue is what "I ought to do" and not whether "I can do it" (Fischer, "Recent Work on Moral Responsibility"). It may seem that the type of argument one can put in support of the *W*-principle is the one articulated by van Inwagen and therefore, the conclusions of the analysis below will apply too.

Carleton University Library

sible for Ridley's death. This indicates that there are situations (such as causal overdetermination)[8] where the association between responsibility and the no-matter-what principle does not hold and thus that the intuition behind this principle can be resisted.[9]

The rejection of PAP is also consistent with the views of some libertarians who concede that, while responsibility can be upheld in Frankfurt scenarios in the absence of robust alternative possibilities, the lack of determinism does rule out responsibility because it does not allow the agent to be the true *source* of the action. As formulated by Michael McKenna,[10] "Source incompatibilists hold that determinism does rule out free will. But it does so, not because it rules out alternative possibilities, but instead, because, if true, the sources of an agent's actions do not originate in the agent but are traceable to factors outside her" (*ibid.*, p. 201). A related position has been taken by Kane, who concedes that some responsibility-bearing actions are determined by antecedent mental states (or willings), but insists that some indeterminism in the causal chain that precedes an action is necessary for the agent to possess *ultimate responsibility* (UR) for that action. This is because he fears that otherwise, the determinism would shift the source of the explanation of the action beyond the agent: "if these willings were in turn caused by something else, so that the explanatory claims could be traced back further to heredity or environment, to God, or fate, then the ultimacy would not lie with the agents but with something else" (*op. cit.*, p. 4).

---

[8] In cases of overdetermination, one may allow "sum events" (the two shootings) as causes in order to justify the intuition that each of the agents has partial responsibility for the death of Ridley. This, however, does not detract from the point that the no-matter-what principle, as applied to each agent alone, is invalid.

[9] The same logic can be used to develop a Frankfurt-type scenario that eliminates the need for a sign predictive of the choice. Consider for example, the following variation on a scenario proposed by Mele and Robb. The agent is faced with a choice between two options, *A* and *B*. Unknown to the agent, the intervener initiates in the agent's brain a deterministic process, *D*, parallel and independent from the indeterministic process, *L* (for Libertarian), assumed to produce a free-choice in the libertarian sense. The *D*-process hits the *A* choice-node (the intervener wants the agent to choose *A*) and the *L*-process hits the *A* or the *B* node (indeterministically) at precisely the same moment, *t*. Either of them (in isolation) is sufficient to cause a decision; however, *D* is stronger, so that in case of divergence its effect prevails. In the case that both processes hit the *A* unit, we seem to be in the situation where the agent chooses *A* and is responsible for it, despite the fact that she could not have *chosen* otherwise (as the *D* process would have overwritten the *L* process). One objection to this could be based on the no-matter-what principle. The agent could claim to be absolved of responsibility since, even if his *L*-process chose otherwise, the same action would have been performed. This case, however, is rather similar to scenario (S_2) above, where two agents shoot (and kill) a person independently. It thus seems that the agent at least shares responsibility for her action despite the fact that she could not have chosen otherwise.

[10] McKenna, "Robustness, Control and the Demand for Morally Significant Alternatives: Frankfurt Examples with Oodles and Oodles of Alternatives," in McKenna and Widerker, eds., *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Burlington, VT: Ashgate, 2003), pp. 201–17.

In order to further support the necessity of indeterminism for UR, Kane gives his *manipulation* argument. According to Kane, we think that an agent, whose actions are the outcome of CNC control by another agent, lacks autonomy and control and UR over her actions, even when those are caused by her own (induced) motivations. Kane uses the illustration of B.F. Skinner's Walden Two, a Utopian community where peoples' values are engineered by behavioral scientists without inducing any sense of coercion. According to Kane, the Walden Two members lack autonomy and UR over their actions, which issue from motivations induced by the indoctrinator. Yet, as Kane argues, it is not clear in what way determination by physical law differs from CNC control, since in both cases the agent's actions appear to be determined (thus lacking alternative possibilities) and beyond her control. According to Kane, only an indeterministic process taking place during character formation, from which the action issues, can salvage UR (*op. cit.*, p. 4).

There are two problems with these arguments. First, the concept of UR has been argued to be logically incoherent and, second, it is not clear that indeterminism can help, as it runs up against the luck counterargument.[11] The former point has been compellingly articulated by Galen Strawson,[12] who argued that an agent acts the way she acts because of the way she *is*, and thus to have UR, the agent needs to be responsible for the way she is. This leads to infinite regression or to a controversial notion of self-causation. Although indeterminism in the choice mechanism may allow the agent to act in a different way (and perhaps form a different character) from the way she actually does, this solution is vulnerable to the luck objection. As argued by Mele, "If there is nothing about the agent's powers, capacities, states of mind, moral character, and the like that explains the difference in outcome, then this difference is just a matter of luck."[13] Notice, however, that the argument from luck, can be deployed even against compatibilistic theories, since inborn biological factors or environmental influences during infancy (that influence the forming character) are also beyond the agent's power.[14]

---

[11] Good presentations of this argument appear in: Clarke, "Free Choice, Effort and Wanting More," *Philosophical Explorations*, II (1999): 20–41; Mele, "Review of *The Significance of Free Will* by Robert Kane," this JOURNAL, XCV, 11 (November 1998): 581–84, and his "Ultimate Responsibility and Dumb Luck," *Social Philosophy and Policy*, XVI (1999): 274–93; van Inwagen, "Free Will Remains a Mystery," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 1–20; but for a counterargument, see Mark Balaguer, "A Coherent, Naturalistic, and Plausible Formulation of Libertarian Free Will," *Noûs*, XXXVIII, 3 (2004): 379–406.

[12] Strawson, "The Unhelpfulness of Indeterminism," *Philosophy and Phenomenological Research*, LX, 1 (January 2000): 149–55.

I believe that a strong conception of UR may, indeed, be incoherent. This issue is controversial,[15] and I will not insist on it here. I will argue instead for a weaker version of ultimate responsibility (UR*), which does not involve self-causation. Rather, UR* requires that the action is one that the agent produces by exerting a certain type of control (to be elaborated below) and which, in addition, is not uniquely determined by anything[16] besides or antecedent to the agent. This prevents Kane's concern about the shift of responsibility beyond the agent, and I will argue that it is not blocked by determinism. Central to the theory I deploy is the acceptance of *constitutive luck*, which should help us resist the idea that heredity or influences of early infancy undermine responsibility. As eloquently discussed by Nicholas Rescher, "identity must precede luck" (*op. cit.* p. 157); we are not born as "bare particulars" that are later allocated with properties. Rather, the properties we are born with are part of our identity and thus are the basis for attributions of responsibility and desert. In section IV, I will argue that this concept of constitutive luck applies also to indeterministic events that affect moral choices of the type considered by libertarians.

Turning now to the manipulation argument, there are two strategies available to compatibilists. The first one, *hard compatibilism*, is to deny the claim that CNC manipulation necessarily eliminates responsibility.[17] As argued by David Blumenfeld (*ibid.*), the most persuasive manipulation scenarios are the ones where the agent's personality is suspended by the controller, who temporarily imposes his motivational structure on the agent. Hard compatibilists admit that, in such situations, when the agent recovers (her previous personality), she is not to blame for actions she did during the interval she was under the control of the manipulation. They claim, however, that a manipulated agent whose value system is transformed forever, and who (due to the manipulation) fully identifies with the evil acts she does as a result of her new (and now persistent) motivational states, is responsible (and can be blamed) for her acts. The second strategy, *soft compatibilism*, concedes that responsibility is undermined even in manipulation scenarios of the latter form.[18] I believe that which of these two strategies we endorse hinges to a large degree on whether we think that the manipulated agent and the pre-manipulated one share the same personal identity.[19] Other cases of manipulation, such as the indoctrination of young children, where a pre-manipulated agent did not exist, trigger more ambiguous intuitions.[20]

---

[13] Mele, "Review of *The Significance of Free Will* by Robert Kane," p. 583.

[14] See Smilansky.

[15] See Balaguer for arguments that indeterminism in torn decisions does not undermine control and authorship of actions.

[16] Excluding "the state of the whole universe," which is a noninformative *source of explanation.*

[17] See, for example, Frankfurt, *The Importance of What We Care About* (New York: Cambridge, 1988); David Blumenfeld, "Freedom and Mind Control," *American Philosophical Quarterly*, XXV, 3 (July 1988): 215–28; Tomis Kapitan, "Autonomy and Manipulated Freedom," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 81–103.

In light of these conflicting intuitions, I take the manipulation argument for PAP to be indecisive. Nevertheless, I believe that this argument does pose a serious challenge to compatibilism. Even if responsibility is not undermined by manipulation, as hard compatibilists hold, another feature of a free agent, her autonomy, does seem to be undermined in CNC manipulation cases. Agents strongly resent to having their autonomy violated,[21] and this is an important property of free agency that needs to be accounted for. The challenge is to account for the difference between CNC control and causal determination of action by physical law, with regard to the agent's autonomy. This requires an account that does not involve indeterminism and can nevertheless distinguish between determination and compulsion.[22]

---

[18] See Mele, *Autonomous Agents*; Ishtiyaque Haji, *Deontic Morality of Control* (New York: Cambridge, 2000). Mele provides a number of colorful illustrations of CNC manipulation. One such example involves a young philosopher, Beth, whose values and pro-attitudes are modified (at the request of her dean) by "new-wave" engineers to match those of Ann (a more industrious philosopher in the same department). According to Mele, after the modification, although Beth and Ann became psychological twins, Ann is autonomous (and responsible for her actions) while Beth is not.

[19] Both Mele and Haji, for example, hold that personal identity is maintained under manipulation.

[20] Consider, for example, Haji's scenario, of a monk raised at the Franciscan order, and who fully endorses (and is happy with) the value system he was indoctrinated with. I believe that whether the monk is responsible (or not) for good actions that follow from his engineered values is debatable.

[21] Consider the following case of manipulation, discussed by Patricia Greenspan ("The Problem with Manipulation," *American Philosophical Quarterly*, XL (2003): 155–64). In this scenario, the students in a psychology class have conditioned their behaviorist professor to move out from his preferred corner of the room by manifesting various signs of comprehension and of attentiveness. In this case, although the professor makes a free and uncoerced choice, his decision is manipulated to fit someone else's end and thus his autonomy is violated.

To conclude, thus far: in the absence of a decisive positive argument favoring PAP,[23] I believe that a theory of free will and responsibility should not assume that, for any action the agent is responsible for, there has to be an alternative possibility that is accessible from the state of the world that preceded the action. (Note, that rejecting PAP does not rule out indeterminism playing a role in our being able to justify the commonsense intuition that we are *initiators* of actions, as suggested by some *soft libertarian* theories.[24])

The theory I propose here is therefore not constrained by the need to respect PAP and is thus compatibilist about determinism and moral responsibility. Nevertheless it does not presume the truth of determinism. Although it may be impossible to establish metaphysical claims, solely from scientific evidence and theory, I believe that quantum mechanics makes it not unlikely that the laws of physics are indeterministic. While it is generally accepted that compatibilism does not depend on the truth of determinism (and that it only insists that determinism does not undermine free will and responsibility), the task of showing that it is robust enough to account for control and responsibility in a world where the relation between the mental states of agents and their action is probabilistic has not been much addressed. The aim of this paper is to do so, while at the same time providing a compatibilist distinction between cases of physical determination and compulsion.

II. DOXASTIC FREEDOM AND PSYCHOLOGICAL OVERDETERMINATION

A compatibilistic theory that does not respect PAP needs, nevertheless, to respect some weaker "freedom" principles. Frankfurt's original suggestion was that moral responsibility does not require alternative possibilities, but that it requires, nevertheless, that the agent has not performed her action *only* because she could not do otherwise. I rely here on a particular interpretation of this idea presented by Robert Cummins,[25] according to which if an agent believes that a desire is irresistible and this belief is part of the reason why she performed the action, she is not responsible for so acting. But, if the fact that the agent *could not do otherwise* did not play any role, as a psychological variable, in the deliberation, the agent is responsible for her action. Accordingly, responsibility is compatible with physical determinism, since an agent "may not know whether determinism is true, and even if she knows, she need not know which acts are determined, and thus the lack of an open alternative plays no role in her deliberation" (*ibid.*, pp. 412–13).

---

[22] Obviously such an account cannot rely on the distinction between determination by another agent versus determination without such an agent; for example, in *Autonomous Agents*, Mele, who offers a compatibilist account of autonomy, agrees that if an agent's values are modified as a result of natural forces (say due to strange electromagnetic fields during to a trip to the Bermuda triangle) the agent is no more autonomous (and deserving of blame/praise for actions performed under those modified values) than an agent whose values were explicitly set up by evil value engineers.

[23] Note also that contra the libertarian intuitions for PAP there are equally compelling intuitions in support of the idea that we would be responsible for actions even in a deterministic world. In his "Recent Work on Moral Responsibility," Fischer has argued that if we were to find out, from a panel of physicists, that the laws of nature are ultimately deterministic, we would not feel that all basis of responsibility attribution and moral justice has evaporated.

[24] See Mele, "Soft Libertarianism and Frankfurt-style Scenarios"; Clarke, "Modest Libertarianism," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 21–45.

In line with this analysis, I assume that an agent is responsible for an action only if knowledge or belief in lack of alternative possibilities is not a part of the deliberation that leads to the action. For example, we tend to absolve of responsibility a guard who does not oppose an armed robbery because he incorrectly believed (due to good reasons) that his pistol was empty of ammunition. Although, in reality the guard had an alternative possible action, he believed he did not have one, and this was the decisive factor in his not opposing the robbery.[26]

This doxastic condition (the agent believing an alternative option to exist), is too weak for responsibility, as it is satisfied in cases of psychological overdetermination, such as addiction and phobias.[27] In order to rule out responsibility in such situations, within a compatibilist framework, John Fischer[28] has introduced the requirement of reason sensitivity. Accordingly, while responsibility does not require the ability to do otherwise (in an identical situation), it does require sensitivity to the input of reasons: the agent should be able to do otherwise, if some (additional) "sufficient" reasons are provided.[29]

---

[25] Cummins, "Could Have Done Otherwise," *The Personalist*, LX, 4 (October 1979): 411–14.

[26] The importance of epistemological cognitive requirements of responsibility has also been emphasized by Carl Ginet ("The Epistemic Requirements for Moral Responsibility," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 267–77), and by Kapitan ("Doxastic Freedom: A Compatibilistic Alternative," *American Philosophical Quarterly*, XXVI, 1 (January 1989): 31–42). Kapitan formally developed a principle first suggested by Daniel Dennett (*Elbow Room* (New York: Oxford, 1984)), according to which the relevant sense of *can* for responsibility is relative to the knowledge of the agent; "something is epistemically possible for Jones if it is consistent with everything Jones already knows," epistemic possibility supplying the "useful notion of can" (p. 184). Kapitan furthermore formulated two conditions that are necessary for a free action (in the doxastic sense): *efficiency* and *contingency*. The former involves the agent's presumption that he would *A*, were he to choose to *A* (and the converse). The latter involves the agent's presumption that the choice is as yet contingent (where the contingency is evaluated relative to the set of cognitive states held by the agent during the choice).

[27] Psychological overdetermination is related to actions performed under "irresistible" desires; see Mele, *Springs of Action* (New York: Oxford, 1992), chapter 5, for further discussion.

To conclude, I will assume that a theory of free will and responsibility needs to allow agents two types of freedom: doxastic ability to do otherwise, and reason sensitivity. In addition, since an agent can only be responsible for *intentional* actions that she controls, the theory needs to account for intentional control. The next section presents the theory, focusing on the key aspect of teleological control and its relation with determinism, indeterminism, and stable behavioral patterns that underlie the predictability of events.

III. TELEOLOGICAL GUIDANCE CONTROL AND PREDICTABILITY

Intentional explanation of actions typically assume that actions are made for reasons and that having such reasons (more precisely, a combination of desires and relevant beliefs) is a causal factor in the generation of actions.[30] For example, we may explain the intentional action of an agent going into the kitchen, as being *caused* by her desire for coffee and her belief that she can find it in the kitchen. Yet, it seems that there is an important aspect of intentional explanation of actions that is not automatically captured in a crude causal account, and this is their teleological or goal-oriented character.[31] Thus we perceive the action of going into kitchen as directed towards a goal (for example, "the agent went into the kitchen *in order to* get coffee"). Here I contend that any intentional action is teleological (and goal oriented), even if its goal does not extend beyond the action itself (say "raising one's arm," or "driving a car on a specified trajectory") and that a necessary condition for such actions is the deployment of a type of control, which I label *teleological guidance-control* (TGC) and which I argue to be necessary for responsibility.[32] I start with an informal and intuitive grasp of what TGC is, using an example from Tomis Kapitan.[33] Having done so, I move to present a substantive account of the physical features of this type of control and its relation with causal laws and their properties, such as *predictability, determinism,* and *indeterminism.*

---

[28] Fischer, *The Metaphysics of Free Will.*

[29] See Fischer and Mark Ravizza, *Responsibility and Control* (New York: Cambridge, 1998) for a more refined version of this account, which states that "an agent is morally responsible for an action insofar as it issues from his own, moderately reason-responsive mechanism" (p. 86), and A.C. MacIntyre, "Determinism," *Mind*, LXVI (1957): 28–41, for an early similar proposal. Note also that even in typical cases of phobia, one may be responsive to some extreme/exceptional reasons: throwing flames into an agoraphobic's house, could persuade the latter to finally leave the place (Mele, "Soft Libertarianism and Frankfurt-style Scenarios"). The input sensitivity thus needs to take place not only under such "exceptional" situations.

[30] Donald Davidson, "Actions, Reasons, and Causes," this JOURNAL, LX, 23 (November 21, 1963): 685–700.

[31] See, for example, George M. Wilson, *The Intentionality of Human Action* (Stanford: University Press, 1989); Scott R. Sehon, "Teleology and the Nature of Mental States," *American Philosophical Quarterly*, XXXI, 1 (January 1994): 63–72, and Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations," *Pacific Philosophical Quarterly*, LXXVIII (1997): 195–213.

Consider the following scenario presented by Kapitan.

> (S3) "...the pilot and the co-pilots of a passenger plane suddenly die, en route, due to a poison ingested before takeoff. The head steward, apprised of the dreadful situation, is faced with the task of guiding the plane to a safe landing. He knows nothing about flying the plane, but were he to press certain buttons and levers, and manipulate the steering mechanism in certain way—actions he is able to perform—the plane would land safely on the designated runway. As it is, he fiddles madly with the controls and manages to do something that results in the plane landing, though, unfortunately, not safely. All aboard perished, except for the steward himself, who survived with minor injuries. Should he be blamed for not bringing it about that the plane landed safely? Was he responsible for bringing it about that it landed in a way that all the passengers were killed?" (*ibid.*, p. 423).

In his analysis of this situation, Kapitan locates an important condition for an action to be under agent control. The way in which the action is issued needs to be *reliably* determined (or ensured) by what the agent has done. While the pilot has *guidance* control over the landing of the plane (she was able to ensure its landing), the steward did not (even if by luck he stumbled over the correct sequence of movements that landed the plane safely). Notice here, the key requirement of *reliable* determination. Plain determination is not enough. If the world is deterministic, both the steward and the pilot landings are determined (by the laws of nature and the state of the world). But how do we distinguish reliable from plain determination?

One way to cast this issue would be in terms of *predictability* (I will address the underlying metaphysics in a moment). We could say that for an event (or state of affairs) to be *reliably* determined by the state of a system at a given moment, its happening needs to be predictable

---

[32] The term "guidance control" is also used by Fischer, but in a different sense (see the section on "reason sensitivity."

[33] Kapitan, "Modal Principles in the Metaphysics of Free Will," in J. Tomberlin, ed., *Philosophical Perspectives: Metaphysics,* x (1996): 419–45.

(in some sense) from the state of the system at that moment. Such an idea was used by Daniel Dennett[34] in his *intentional stance* theory, in which we attribute intentional states to systems (or organisms) whose behavior shows a reliably predictable pattern (which is stronger than what can be predicted from Newtonian equations and is not derivable à la Laplace). Animals, plants, and even lightning bolts and thermostats, are relevant examples. An important feature of the kind of predictability in question is its being robust given variations (as illustrated in the pilot scenario). Characterizing control and responsibility in terms of this kind of predictability, however, has the disadvantage that it seems to make control and responsibility relative to one's scientific skills. In response to this worry, Dennett argued that intentional predictability is grounded on the fact that there are objective patterns that "impose themselves, not quite inexorably, but with great vigor, absorbing physical perturbation and variation" (*ibid.*, p. 27). What the nature of these objective patterns is, however, was not explained.

A main aim of this paper is to account for the objective structure that mediates the patterns enabling domains with predictability in the word of physical objects and organisms. The upshot, I argue, is that this requires a *teleological* system with goal-directed behavior, which will reach the same end state in face of perturbations. As many critics have noticed, however, "there is a prima facie tension between the common sense account of ourselves as agents and the scientific view of human beings as physical objects. Notions like action and goal-direction appear to have no role in purely physical descriptions of the world. Planets, rocks and elementary particles, do not *do* things; if we are no different, in principle, than these things, then our status as agents who do things can be legitimately put into question."[35]

I endorse a causal theory according to which teleological behavioral patterns arise within a causal dynamical system. Such a theory, therefore, needs to account for the apparent tension between causal and teleological behavior, which is probably due to the fact that there appears to be no proper place for teleology in the physical laws of nature. A central principle of both deterministic (Newtonian mechanics and electromagnetism) and indeterministic (quantum mechanics) physical theories is that the change in the state of a system (and thus its future) depends only on its present state. Both theories have also fully reversible dynamics; no place is thus left for "tendencies" to reach an end state, as was the case in Aristotelian physics. It is perhaps for these reasons, that teleological systems are most often characterized in relation to a "design" or a function.[36] Accordingly, an artifact is teleological because of its doing what it was designed to do; we and other biological organisms are teleological because of our doing what the Darwinian evolution "designed" us to do. While I agree that evolution helps to explain how teleological systems (of biological kind) emerged, I believe that their proper characterization should be independent of the evolutionary origin.[37] Here I propose a different account of teleological systems, possessing guidance control (TGC) and consistent with causal laws.[38]

---

[34] Dennett, *The Intentional Stance* (Cambridge: MIT, 1987).

[35] Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations," p. 197; see also Nagel, pp. 113–19, for a similar discussion.

Here is the main idea. Teleological systems require a property that is stronger than determinism and which involves a *counterfactual* type of determination, called an *attractor*. TGC systems (or agents) generate stable behavioral patterns, in which events (or states of the world) are being determined by a goal state of the system (agent), in a set of possible worlds similar to (and including) the actual one; the fact that the event is determined to take place in a set of counterfactual situations (similar to the actual one) reflects the requirement of being *reliably* determined. Critically, neither determinism nor indeterminism is sufficient or necessary for TGC, which is typically manifested at a macro-level of description, involving collective states. I explain these ideas below, starting with the conditions that enable the appearance of teleological goal-directed systems.

Consider, first, the mixing of a liquid between two connected containers, originally filled with hot and cold liquid, respectively. Under the second law of thermodynamics, an irreversible process of flow takes place between the containers, equalizing the temperatures. Although the behavior of individual molecules is subject to reversible causal laws (and in practice, contra Laplace, unpredictable), the temperature of the liquid in the two containers, a collective (or macro) state variable, shows a teleological pattern: it *tends* toward a state of equilibrium. The teleology of this equilibrium state is due to its finality or *convergence* type of behavior; the macrostate *converges* toward the equilibrium, since it ends there even if perturbations are applied trying to change its course. (Other examples of collective macrostates displaying teleological behavior are the vortex of a whirlpool that sinks whatever small objects get close to it, or a lightning bolt that is attracted to the best object that conducts electricity in the ground. In both cases, although the trajectory of individual molecules is subject to causal physical laws (but is in practice unpredictable due to chaotic dynamics), the collective state that corresponds to the vortex (its center and angular momentum) or to the lightning (its center and momentum), displays a stable and predictable behavior. An end-state with this property is called an *attractor*. In an attractor dynamics, the "space of possibilities" shrinks as trajectories converge towards the attractor (see Figure 1; more discussion of this figure follows in section IV), leading to a decrease in *entropy*.[39] Crucially, the property of convergent or attractor dynamics is orthogonal to the determinism/indeterminism distinction—one can have determinism with or without convergence; and one can have indeterminism with or without convergence.[40] The role of attractor states in guidance control of actions is illustrated in the following example.

---

[36] Dennett, *The Intentional Stance*; Ruth Millikan, "On Swampkinds," *Mind and Language*, XI (1996): 103–17.

[37] The over-reliance of theories of content on evolutionary teleology leads to some counterintuitive consequences. For example, swampmen (creatures that are accidentally created out of molecules in a swamp, with body and brain structures identical to ours, but without our evolutionary theory) have no beliefs and desires (see Millikan). Liberating teleology from its link to evolutionary design opens the way for swampmen liberation (Louise Antony, "Equal Rights for Swamp-persons," *Mind and Language*, XI (1996): 70–75; Usher, "Comment on Ryder's SINBAD Neurosemantics: Is Teleofunction Isomorphism the Way to Understand Representations?" *Mind and Language*, XIX (2004): 241–48).

[38] See also Mele, "Goal Directed Action: Teleological Explanation, Causal Theories, and Deviance," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 279–300.

Consider a simple goal-directed system: a teleguided rocket. The mechanism that enables the rocket to display teleological behavior by "pursuing" the target (seen here as its goal) is based on a feedback loop and an error correction mechanism. This mechanism stabilizes the rocket's trajectory towards the target despite external perturba-

---

[39] The entropy (or the Shannon information) reflects the degree of uncertainty of the system (or the number of possible microscopic configurations, consistent with the macroscopic description (for example, temperature). Under the second law of thermodynamics, the entropy increases. This is the case for the mixing of liquids, but not in the other two examples (see next footnote).

[40] The deterministic Newtonian/Hamiltonian dynamics satisfy the Liouville theorem, according to which the volume of an area in the "phase space" (the space of all possibilities of the variables that define the system) remains constant as the trajectories evolve, and a similar property exists in quantum mechanics. In an attractor dynamics, the volume in the configuration space shrinks as the trajectories converge, producing a decrease in entropy. In either, a Newtonian or a quantum world, it is possible, however, that a subsystem (corresponding to a collective variable, for example, the tornado vortex) behaves according to attractor dynamics provided that the rest of the system, which involves other variables, compensates for it (by increasing entropy). Collective-state variables that control actions in the brain correspond to population codes or cell assemblies, whose behavior is characterized by attractor dynamics (see Figure 1). Other variables that are irrelevant to the controlling behavior (for example, heat released by the body) involve a divergence dynamics compensating for the decrease in entropy.

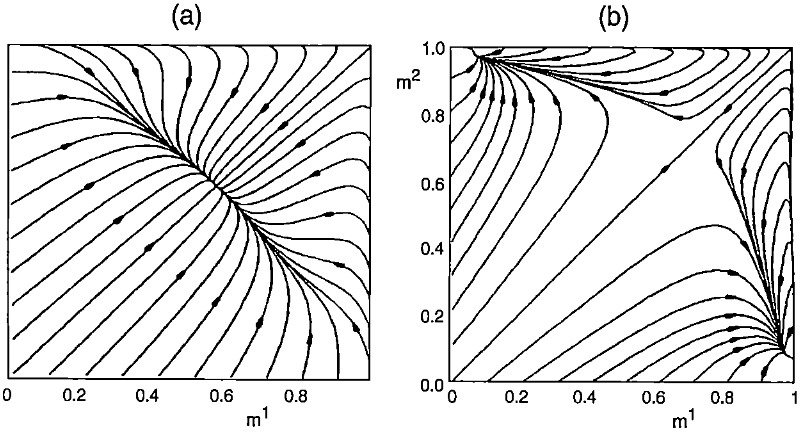


Figure 1: Attractor dynamics in a deterministic nonlinear neural system. The dynamics are indicated by the arrow flow, showing convergence to attractor points. (a) The single attractor corresponds to a scenario where all the trajectories converge at the same end point, corresponding to a case of overdetermination. If the various starting points correspond to informational-relevant inputs (reasons or data in favor of one or the other of the options) for a specific decision, the single attractor implies a total lack of input sensitivity; one cannot not do otherwise, even with different input relevant to the situation. (b) A multiple-attractor scenario, where although the trajectory is bound to end up in one of the two attractors, there are bifurcation points (input-relevant situations) where a different attractor is reached. Each attractor has a *basin of attraction*—a nonzero volume of states that map into that attractor. This ensures that input sensitivity is not too weak, holding not only for the most exceptional situations (see footnote 29).[41]

tions. If a gust of wind makes the rocket deviate from its planned trajectory, the correction mechanism will bring it closer to it. This confers a teleological behavior on the rocket: it has a type of purpose or finality. When the rocket is launched repeatedly at the same target, the actual trajectory and the forces applied may vary, but the convergence upon the target prevails. *Stability under perturbations* is a property that distinguishes the rocket from nonteleological type of projectiles, which may happen to hit upon a target but do not *converge* on it. As with the water temperature, the rocket mechanism is characterized by an attractor. This means that it reaches its target not only in the world corresponding to the actual situation, but also in similar possible worlds (say, when its trajectory is subject to wind perturbations or when the target attempts to escape). This makes such control mechanisms highly effective in achieving their goal, and although their action is purely causal, they generate the appearance of a teleological, purpose-oriented process.

---

[41] Reproduced from D. Horn and Usher, "Dynamics of Excitatory-inhibitory Networks," *International Journal of Neural Systems,* I (1990): 249–57.

The need for a correction-stabilization mechanism, in the action control of organisms, is obvious in an indeterministic world. This need exists, however, even if determinism holds. Since agents cannot have full knowledge of all the environmental inputs, a certain degree of uncertainty and randomness needs to be accepted as an inevitable feature of the environment.[42] Therefore, the problem that an agent faces is not how to get from *A* at t_1 to *B* at t_2 given the precise state of the world at t_1, but rather how to get from *A* at t_1 to *B* at t_2 given the (limited) information possessed and in spite of a variety of unexpected interventions or perturbations during the interval. This requires a property that is stronger than determinism and involves a counterfactual type of determination. I propose that this property is necessary for the intentional control (of the TGC type) that is used in the situation described by Kapitan. It is in virtue of being able to guide the plane to landing in face of a variety of counterfactual situations (wind changes, other approaching planes, and so on) that the pilot (but not the steward, even if by luck he managed a safe landing) is in control (has TGC) of the plane.[43]

One aspect of TGC is crucial. TGC is not only a property of a situation (in terms of an attractor). When TGC exists one can also locate a *center* that generates it and which is responsible for the entropy reduction. This center is the collective state that governs the attractor (for example, the error-correction mechanism for the rocket, or the vortex center and its rotation for the whirlpool). For an agent to be in control of her action, it is critical that the TGC is produced by an intentional state (for example, an intention) of the agent. If, on the other hand, the TGC is *externally* produced, then the agent is not controlling her action but, rather, she is being controlled. Being externally produced means that the attractor variable is not an item within the mental life of the agent. The obvious case is when the controlling variable is physically external to the agent (say, as when a person is caught in a whirlpool and cannot escape it). In other cases, the controlling variable may be physically internal to the agent, but external to her mental states (for example, when the agent's emotional state is uniquely determined by an ingested substance or when a brain tumor makes the agent behave inconsistently with her own beliefs and values[44]). This will be discussed in section V in relation to manipulation scenarios.

---

[42] See Rescher.

[43] Control can be detected by a decrease in the entropy (or increase in Shannon information) of the system. The control mechanism of the rocket ensures that a large set of possible configurations at t_1 is mapped into a small set of configurations (or a single state) at t_2. Such entropy reducing (or information generation) mechanisms are therefore quite peculiar and provide a sensitive method to characterize the notion of teleological control and to distinguish it from causal determination.

An important question is: If TGC needs to be internally produced, why is internal production not alone sufficient? Suppose that an agent intentionally raises her arm. In what sense, does one need counterfactual control (in order to ensure that she reaches the same end state under a variety of conditions) when we are dealing with a token case? Moreover, if in other circumstances, the agent fails to raise her arm, does that mean that she did not act freely or intentionally in the present case?[45] This question parallels the question of whether causal determination is sufficient to account for intentional actions, or whether a stronger teleological requirement is needed.[46]

Consider the following two situations.

> (S_4) *Intentional action of a highly constrained agent.* The agent is highly constrained in her action abilities: all she can do is press (or not press) a button, which will trigger a mechanism shooting a person at a prespecified time (the timing is programmed in advance and the agent does not control it). The agent presses the button and kills the person.

I contend that, for the shooting action to be intentional it must be the case that, if the agent was informed (and she trusted the information) that the triggering mechanism had been reversed and that pressing the button will now block the bullet that would have been otherwise shot (at the prespecified time), then she would have refrained from pushing the button. Accepting this, however, indicates that even under the most limited conditions (with a repertoire of only two ways to achieve an action), how the agent would behave in a perturbed condition has a bearing on whether her present action is intentional (note that the action is the *shooting* and not the *pressing the button*). To qualify as intentional, the action needs to have been generated out in a teleological way, so that it can be achieved under some perturbed situations.[47]

---

[44] One may wonder if environmental factors, such as warm weather or dopamine-rich diets may involve a type of external environmental control (I thank the editors for raising this issue). I believe that it is important here to distinguish between those factors that have a *biasing* effect and those which uniquely determine their effect. For example, if warm weather makes people less enterprising or energetic, but it allows for variation in the energy profile among individuals, then it is only one of many factors, which together, contribute to the agent's character (weather has a biasing effect). If, one the other hand, the factor has a very specific effect and it eliminates individual differences (everyone who ingested the substance develops the same character or set of values), then it can be seen as a form of external control by the environment. In practice, the likelihood of this occurring (except in situations, such as the ingestion of a substance triggering uncontrollable aggression, which totally wipe out the agent's control due to lack of reason sensitivity) is vanishingly small without a teleological mechanism.

[45] I wish to thank the editors for raising this question.

[46] Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations."

> (S_5) *Deviant causal chain.* The agent intends to knock over her glass of water in order to distract her debate opponent. However, her intention upsets her and makes her feel so nervous that her hand shakes uncontrollably, striking the glass and knocking it to the floor.[48]

In this case, although the action of dropping the glass is caused by the intention to drop it, no counterfactual reliability exists, and therefore it is not an intentional action. It is precisely because nervousness diminishes TGC, that it cannot *reliably guide* the same action to take place in perturbed situations.[49] I next argue that the scheme for control and responsibility that I presented is robust enough to function even in an indeterministic world and that a succession of attractor and bifurcation stages is essential for responsibility.

IV. BIFURCATIONS, CHOICE, REASON SENSITIVITY, AND NEURAL DYNAMICS

The property of attractor dynamics does not apply constantly over time. (If it did, people would be much more predictable, and boring, than they are!) Rather, one typically finds a process that involves successive stages of attraction (where perturbations are absorbed) and bifurcations (where they are amplified). While the attraction stages correspond to goal states (intentions for action or stable character traits) that reliably predict (and guide) behavior, at bifurcation points the agent's behavior is less predictable and shows a higher sensitivity to the input of reasons (in Fischer's sense, that new evidence would modify the choice). Such bifurcations can appear in two important ways in the genesis of free actions. First, they take place when the agent is faced with a choice between conflicting courses of actions, each of which is supported by some reasons the agent endorses. Second, bifurcation processes take place in the process of acquiring moral values (this is a slower process, related to character formation, where conflicting sources of information are interpreted).

---

[47] An intentional action involves some knowledge condition of the goal; note that a mechanism of error correction (such as that of the guided rocket) needs to possess a *representation* of the target and a procedure of action toward it.

[48] This is adapted from Mele, *Springs of Action*; Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations."

[49] See further discussion, in Sehon, "Teleology and the Nature of Mental States," and "Deviant Causal Chains and the Irreducibility of Teleological Explanations."

One system with both attractor and bifurcation dynamics is a multi-attractor system. Figure 1 illustrates such a system, using computational studies of neural cell assemblies. The requirement of multiple attractor dynamics is important for ruling out responsibility in cases of psychological overdetermination and is consistent with Fischer's proposal that responsibility requires sensitivity to the input of reasons. (Notice that I do not require that every decision starts from a bifurcation point, but only that an alternative attractor exists.) Research in neural networks has shown that such multi-attractor dynamics can also be *intermittent*, with time intervals of stability (convergence towards an attractor) followed by intervals of instability and bifurcations, where the convergence dynamics is switched into one of divergence, before a new attractor state is reached.[50] Moreover, neural systems can form new attractor states as a result of self-organization.[51]

It is important to note that such bifurcations, at the level of mental (psychological) states, are possible even if physical determinism holds. Consider the situation where at time t_1 the mental macrostate of the agent is close to the boundary between the two choice attractors in Figure 1b. Even if *physical* determinism is true (all tokens of the same type of brain state, B_1, at t_1, evolve into tokens of another type of brain state, B_2, at t_2) psychological determinism does not need to hold (different tokens of the same type of psychological macrostate, M_1, at t_1 may evolve into tokens of multiple types of mental states at t_2–M_2, M_3, and so on). This is due to the fact that there are alternative brain microstates at t_1 (for example, p_1, p_2) that are equally sufficient for the mental macrostate M_1 (at t_1) but, at bifurcations, the dynamics are sensitive to the differences between the microstate tokens. For example, these microstates can deterministically evolve from t_1 to t_2 as: $p_1 \rightarrow q_1$, $p_2 \rightarrow q_2$ with q_1 and q_2 sufficient for different psychological macrostates (M_2 and M_3, respectively).[52]

---

[50] O. Hendin, D. Horn, and Usher, "Chaotic Behavior in an Excitatory-Inhibitory Network," *International Journal of Neural Systems*, I, 4 (1991): 327–35.

[51] See, for example, W. Dong and J.J. Hopfield, "Dynamic Properties of Neural Networks with Adapting Synapses," *Network: Computation in Neural Systems*, III (1992): 267–83.

[52] See also Nick Zangwill, "Daydreams and Anarchy: A Defense of Anomalous Mental Causation," *Philosophy and Phenomenological Research* (in press).

Additional reasons for bifurcation processes in choice or character formation are indeterminism and the presence of "noisy" input or epistemic uncertainty, such as a tune on the radio bringing to mind a memory that biases the agent one way or the other. A simple way to take such factors into account (standardly used in neurocomputational models of choice) is by adding a source of noise on top of the dynamics of the choice macrostate (for example, quantum indeterminism may lead to minute stochastic fluctuations in synaptic transmission events).

Does such open-endedness during deliberation undermine control and responsibility? The first thing to notice is that due to the attractor dynamics that are robust to perturbations, the effect of indeterminism (as well as that of epistemic noise) shows only at attractor boundaries. Therefore the objection that indeterminism diminishes control by making the agent choose or act against her "better judgment"[53] does not apply (in the absence of attractor dynamics, the noise may modify the "better judgment"). At bifurcations, however, the agent's will is divided. If the agent has competing reasons for two courses of action, no matter which of these reasons she decides to act on, the agent has authorship over the action.[54]

In agreement with Kane and Mark Balaguer (*op. cit.*), I believe that indeterminism does not undermine the fact that, when an agent performs an action and this action is probabilistically caused by her intentional states (for reasons consistent with the agent's character, motivation, and so on), the agent bears responsibility for her action. Assume for example that given an agent's character and current situation, the probability of the agent *A*-ing is 0.1 and the probability of her *B*-ing is 0.9. Consider one of the rare cases where the agent *A*-ed. I maintain that as long as there is any probability (larger than zero) for performing an action as a result of a mental state (and for reasons related to it), when the agent performs the action she is fully responsible for doing so. The agent's responsibility for the action is grounded in the fact that she has mental states that are directed towards that action and make it more likely.

Nevertheless, as Kane admits, although probabilistic causation explains the action as a causal outcome of the agent's mental states, it cannot explain why the agent chose *A rather than B.* Since such a *contrastive* explanation is unavailable, the fact that the agent chose *A rather* than *B* remains random. But as I argued above, this does not undermine the responsibility for *A*-ing, if *A*-ing is consistent with the agent's character and motives. If, however, we inquire about the agent's UR—that is, her responsibility for having the character and the motives that she has—I agree with the compatibilist critics that the lack of a contrastive explanation is likely to become problematic: at birth there is little mental repertoire to ground responsibility for indeterministic decisions, and if the developing character depends on them (as Kane requires), the way this character turns out may well be considered to be a matter of luck.[55] As we saw earlier, however, the problem of luck is not exclusive to indeterministic theories; deterministic theories are faced with the same problem, when challenged about hereditary and biological factors that affect an agent's character. The solution I prescribe (and which diverges from Kane's) is to allow that there is *constitutive luck* in both cases. Accordingly, an agent's character is an evolving system, which is affected by (and absorbs) a variety of "random" factors, indeterministic events being only one possible subclass.[56]

---

[53] Mele, *Autonomous Agents,* p. 203.

[54] See Balaguer for further aspects of this argument.

To summarize: indeterminism is not necessary for responsibility and control. However, it also does not undermine them, if it takes place at bifurcations. I contend that some type of TGC mechanism is necessary for any intentional responsibility-bearing action. Such mechanisms correspond to attractors in dynamical systems with variable intervals of persistence (from minutes to years)[57] and they yield to bifurcations, which lead to novel attractors. The responsibility for an action can then be traced back to the relevant attractor that has TGC over it. For example, we may trace back the responsibility for a building's breakdown to the tornado that destroyed it, or to the person who demolished it; crucially we stop here, as we are not able to track the responsibility beyond bifurcation points. Consider a world where no such convergent/divergent dynamics take place. If that world was deterministic (and lacking attractor/bifurcation dynamics), then one could trace the responsibility for an effect all the way to the state of the world at the Big Bang; no other state of the world is "special" in its role at bringing that effect about. By contrast, a world with successions of bifurcations and attractors has special centers of responsibility. In such a world, even if the transition from a bifurcation to an attractor is luck dependent or random, the attractor becomes the ultimate feature that explains why its associated TGC patterns take place, in a way that nothing antecedent to it can do (that is, the attractor has UR*). Even if their origin is grounded in constitutive luck, transient attractors are entities of which it is correct to use the slogan of Harry Truman, "The buck stops here." In the following section I discuss situations where an attractor state was itself TGC-induced by another attractor state; for example, the tornado may have been engineered, or an agent's character created by indoctrination.

---

[55] Bernard Berofsky, "Ultimate Responsibility in a Deterministic World," *Philosophy and Phenomenological Research*, LX, 1 (2000): 135–40.

[56] The theory I propose here is neutral on the question of whether indeterminism, although not needed for responsibility, has a value in allowing agents to "make contributions to the world, that they are not determined to make" (Mele, "Soft Libertarianism and Frankfurt-style Scenarios," p. 135). Whether such indeterministic contributions (or initiations) have a value depends, however, on what we say about the following. Suppose that after a hard open-ended bout of deliberation, we discover that one of the following scenarios is true: (i) the choice was ultimately (in)determined by an irreducible probabilistic factor in the brain choice mechanism (à la Balaguer); (ii) the choice was ultimately determined by a minute external influence (say, a tune accidentally heard on the radio that brought to mind a thought that made us prefer one of the options). Do we really feel that we made a difference to the world in (i) but not in (ii)? I tend towards the view that the degree of authorship and initiation in these two situations is the same, so indeterminism does not matter: it is merely one type of constitutive luck.

[57] Transient attractors correspond not only to mental processes such as intentions for actions or stable character traits, but even to meteorological phenomena such as tornadoes, or social processes such as Communism and Nazism.

V. AUTONOMY AND MANIPULATION

The challenge facing compatibilism is to account for the distinction between determination and compulsion. In the previous section, I used the concept of TGC to account for the distinction between actions that are under agent control and actions that are *merely* determined by physical laws. Actions that are under an agent's control are the outcome of TGC that originates *within* the choice and action production mechanism of the agent. There are situations, however, when the source of the TGC is *external* to the agent. Consider, for example, a manipulated agent. The manipulator may use a variety of techniques (setting up the situation, conditioning, and so on) to make the agent produce a desired action. In a recent article, Patricia Greenspan (*op. cit.*) discusses a variety of manipulation scenarios that violate the agent's autonomy, with the conclusion that one of their central features is that the manipulee is used as a means to the manipulator's end. Note that this involves teleological and counterfactual determination: the manipulator will ensure the agent will produce the action in a variety of nearby possible worlds. This does involve TGC, whose source is *external* to the manipulated agent, as its origin is not part of the agent's mental states. In this situation, the agent is not in control, but rather she is being controlled by the source of TGC. Being under external control that leads to an action is being compelled to do it.

Notice, that being subject to compulsion does not require a cogent agent to do the manipulation, but rather takes place whenever possibilities are reduced (entropy reduction) by an attractor type dynamics whose source is external to the agent. Clearly, other cogent agents are the most likely sources that may exert such entropy-reducing control. However, a noncogent attractor system, such as a tornado vortex or a lighting bolt, could achieve the same.[58]

The central challenge is to distinguish between the "benign" type of influence that occurs in any aspect of normal life (for example, children's education) and the pernicious type (manipulations) that we consider to be autonomy violating and (arguably) responsibility reducing.[59] Compare normal education with the type of manipulation involved in an effective indoctrination (or value engineering) that uses a combination of methods based on positive reinforcement (as described Skinner's Walden Two), manipulation of informational input, and consensus group dynamics.[60] An agent who was subject to such indoctrination can still exert TGC to produce actions that satisfy her goals and motivations, and to this extent she is responsible for these actions; had the agent (or corresponding TGC state) been disabled, the action would not have occurred. In my view, we can agree with hard compatibilists that indoctrinated agents are responsible for their actions. However, I think we can also agree with soft compatibilists and libertarians on two important aspects: First, indoctrinated agents are not autonomous, and second, although they are responsible for their actions, they are not the *last source* of responsibility: this is the indoctrinator who shares the responsibility for his agent's actions.

To further examine the difference between the autonomy of the "normal" and of the indoctrinated agent, consider two agents who exert TGC at present, but who differ profoundly in the way their motivational and belief states were acquired. One acquired her motivational states under normal education, while the other acquired her motivational states under indoctrination. The value systems and motivational states of both agents depend on some external factors. For both, the combination of values that characterizes their mental makeup is determined by the environment. A first difference, one may discern, involves the "complexity" of the *determining function*. For the "normal" agent the determining function involves a myriad of environmental and biological factors including innate (genetic) ones, while, for the indoctrinated agent, the determining function can be specified by a few factors corresponding to the intentions of the indoctrinator at a previous time.[61] Should this matter? Or should one rather say that, after all, determination is determination?

---

[58] Although highly unlikely, a nonTGC event such as the electrical noise in a plane navigation system could neutralize the pilot maneuvers and bring the plane to a safe landing at a different destination than intended. Similarly, an agent may have a uniquely determined change of character (but which maintains her ability for "normal" action control, as does the Beth-character in Mele's autonomy violating scenario; see footnote 18) due to an ingested substance. The likelihood of such autonomy undermining non-TGC events, however, is extremely small.

[59] Kane, *The Significance of Free Will.*

[60] See examples in Greenspan.

Note, however, that there is a more important difference between the two agents, which shows up when we examine counterfactuals. Unlike the normal agent, the mental configuration of the indoctrinated agent is *counterfactually determined* by the indoctrinator. By contrast, the mental configuration that the normal agent ends up with, would diverge significantly given minute variations in environmental input or in biological makeup. The character formation of this agent is the result of a (luck dependent but *constitutive*) self-organizing process, which results into a stable character. Unlike with the indoctrinated agent, the normal agent's character is not *reliably determined* by something besides or antecedent to the agent: she has UR*.

The soft compatibilist can then propose that autonomous agents are those whose TGC mechanisms were not subject to a *fully* determining TGC, whereas manipulated agents' were.[62] Thus, although both agents are responsible for their actions, only the normal agent is autonomous; the indoctrinated agent shares responsibility with the indoctrinator, who is the TGC source of her motivations.

In practice, an indicator of the presence of indoctrination (and loss of autonomy) can be obtained by examining the degree of *diversity* of value configurations of agents raised in a particular environment, and the predictability of their configurations (on the basis of the education program). It is in the nature of a good education program that it allows (or even encourages) diversity and that despite the "best" efforts of the educator, the emerging character of a developing agent is bound to surprise us. On the other hand, highly uniform societies (where differences between agents are minimized) are a strong indicator of the presence of indoctrination and of the loss of autonomy.[63]

---

[61] The determining function maps the state of the world at a previous time, t_1, to the mental state of the agent. With the indoctrinated agent, a time t_1 (when the indoctrinator has formed his plans) exists, where the determining function requires a small number of variables (the intentions of the indoctrinator) and other physical variables are screened out, while with the normal agent, the determining function at t_1 is sensitive to variations in a much larger set of physical variables.

[62] Note that a biasing process, such as warm weather, does not make the biased character insensitive to other factors and thus is not enough to eliminate variation and undermine autonomy. Note also that this is a simplification, since indoctrination (as well as autonomy) is not an all or nothing affair, but rather is a matter of degree.

VI. CONCLUSION

The account of free will and responsibility proposed here shares features with a number of compatibilist and libertarian theories. In particular, it makes use of doxastic freedom and of the reason-sensitivity principles for responsibility. A part of this account, however, was to develop a theory of teleological control, based on attractor/bifurcation dynamics consistent with causal mechanisms, which can function within an environment characterized by uncertainty. Within this theory, intentional control is characterized by a counterfactual determining causal process (teleological guided control, TGC) whose center is a psychological state. Bifurcation processes, where input sensitivity is maximal, taking place in action selection or in character formation, do not undermine responsibility, if one allows for constitutive luck. Crucially, the theory can distinguish between causal determination, which does not undermine autonomy and CNC manipulations, which do. Finally, the theory does not contradict the powerful intuition that processes of deliberation are open ended.

MARIUS USHER

Birkbeck College, University of London

---

[63] A manipulated society, such as that described in Aldous Huxley's *Brave New World* can also "set up" diversity by giving each agent a different indoctrination program. Thus one needs to know more about how diversity was produced in order to conclude that no manipulation was at work.

NEW BOOKS: REPRINTS

ADDAMS, JANE: *Writings on Peace.* New York: Continuum, 2003/2005. xxii, 389 p. Paper $39.95.

ARONSON, RONALD: *Camus and Sartre: The Story of a Friendship and the Quarrel That Ended It.* Chicago: University Press, 2004/2005. x, 291 p. Paper $19.00.

ARRINGTON, ROBERT L., and MARK ADDIS, eds.: *Wittgenstein and Philosophy of Religion.* New York: Routledge, 2001/ 2004. xi, 187 p. Cloth $100.00, paper $32.95.

AYER, A. J.: *Probability and Evidence.* New York: Columbia, 1972/2006. xxviii, 144 p. Cloth $64.50, paper $24.50.

BENJAMIN, WALTER: *Selected Writings, Volume 2, Part 1: 1927–1930.* Cambridge: Harvard, 1999/2005. vii, 463 p. Cloth (entire volume) $47.50, paper $18.95; *Selected Writings, Volume 2, Part 2: 1931–1934.* Cambridge: Harvard, 1999/2005. vii, 460 p. Cloth (entire volume) $47.50, paper $18.95.

BOETHIUS: *Boethius's De topicis differentiis.* ELEONORE STUMP, ed. and trans. Ithaca, NY: Cornell, 1978/2004. 288 p. Paper $26.95.

BROWN, STUART, ed.: *British Philosophy and the Age of Enlightenment: Routledge History of Philosophy, Volume V.* New York: Routledge, 1996/2000/2001/ 2003. xliii, 395 p. Paper $32.95.

CAMP, JOSEPH L.: *Confusion: A Study in the Theory of Knowledge.* Cambridge: Harvard, 2002/2004. x, 246 p. Cloth $42.50, paper $18.95.

CANFIELD, JOHN V.: *Philosophy of Meaning, Knowledge and Value in the Twentieth Century: Routledge History of Philosophy, Volume X.* New York: Routledge, 1997/2003. xxxiii, 466 p. Paper $32.95.

CHOMSKY, NOAM: *Rules and Representations.* New York: Columbia, 1980/2005. lii, 299 p. Cloth $64.50, paper $24.50.

COHEN, JEAN L.: *Regulating Intimacy: A New Legal Paradigm.* Princeton: University Press, 2002/2004. xi, 290 p. Cloth $45.00, paper $19.95.

COLLINGWOOD, R.G.: *An Essay on Philosophical Method.* New York: Oxford, 1965/2005. cxxii, 360 p. Cloth $94.55.

CURD, PATRICIA: *The Legacy of Parmenides.* Las Vegas: Parmenides, 1998/2004. xxix, 280 p. Paper $22.00.

DANTO, ARTHUR C.: *The Philosophical Disenfranchisement of Art.* New York: Columbia, 1986/2005. xxx, 216 p. Cloth $64.50, paper $24.50.

DARWALL, STEPHEN: *Welfare and Rational Care.* Princeton: University Press, 2002/2004. xi, 135 p. Cloth $35.00, paper $16.95.

DASGUPTA, PARTHA: *Human Well-Being and the Natural Environment.* New York: Oxford, 2001/2004. xxiv, 351 p. Cloth $49.95, paper $24.95.

FERRARI, G.R.F.: *City and Soul in Plato's Republic.* Chicago: University Press, 2003/2005. 130 p. Paper $17.00.

FISHER, ALEC: *The Logic of Real Arguments.* New York: Cambridge, 1988/2004. xii, 224 p. Cloth $65.00, paper $22.99.

THE JOURNAL OF PHILOSOPHY

BACK ISSUE SALE

Complete volumes
are now available
at the following discounted prices
in order to reduce our inventory:

Volumes	1 through 98	$10 each
Volumes	99 through 101	$25 each

For more information, contact Pamela Ward,
Business Manager, at pcw1@columbia.edu

Visa, Mastercard, American Express and
PayPal are accepted through our website at
www.journalofphilosophy.org

the review of
metaphysics

a philosophical quarterly

JUNE 2006 | **VOL. LIX, No. 4** | **ISSUE No. 236** | **$15.00**

articles

SARAH BYERS *Life as "Self-Motion" Descartes and the Aristotelians on th Soul as the Life of the Bod*

RICHARD DIEN WINFIELD *Self-Consciousness and Intersubjectivit*

KENNETH WESTPHAL *How Does Kant Prove tha We Perceive, and Not Merely Imagine Physical Objects*

DALIA NASSAR *Immediacy and Mediac in Schleiermacher'* Reden Uber die Religio

DOUGLAS B. RASMUSSEN and DOUGLAS DEN UYL *The Myth of Atomis*

ELIZABETH C. SHAW AND STAFF *Summaries and Comment*

book reviews *philosophical abstracts* *announcemen*

Individual Subscriptions $35.00 Institutional Subscriptions $55.00 Student/Retired Subscriptions $20.

VISA AND MASTERCARD ACCEPTED

Jude P. Dougherty, Editor
The Catholic University of America, Washington D.C. 20064
email: mail@reviewofmetaphysics.org

crítica

REVISTA HISPANOAMERICANA DE FILOSOFÍA

Artículos

MAX A. FREUND
Conceptualismo realista y computabilidad

OLIMPIA LOMBARDI Y MARTÍN LABARCA
Los enfoques de Boltzmann y Gibbs frente al problema de la irreversibilidad

CHRISTIÁN C. CARMAN
La distinción teórico/observacional: ¿favorece o perjudica al realismo científico?

Notas bibliográficas

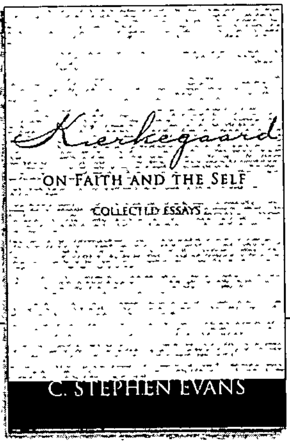
SERGIO F. MARTÍNEZ, *Geografía de las prácticas científicas. Racionalidad, heurística y normatividad*
[J. Francisco Álvarez]

MIGUEL ESPINOZA Y ROBERTO TORRETTI, *Pensar la ciencia. Estudios críticos sobre obras filosóficas (1950–2000)*
[Sandra Lucía Ramírez Sánchez]

Libros recibidos

Vol. 37 / No. 111 / diciembre 2005 / ISSN 0011–1503

Crítica, Revista Hispanoamericana de Filosofía aparece tres veces al año en los meses de abril, agosto y diciembre. Es una publicación del Instituto de Investigaciones Filosóficas (UNAM). Todo tipo de correspondencia deberá ser dirigida al Apartado Postal 70-447, Coyoacán, 04510, México, D.F. Correo electrónico: critica@filosoficas.unam.mx. Versión electrónica: http://critica.filosoficas.unam.mx



Kierkegaard on Faith and the Self

Collected Essays

C. Stephen Evans

University Professor of Philosophy and the Humanities,
Baylor University

A treasure trove from one of the world's finest Kierkegaard scholars ... Like Kierkegaard's own writing, these essays are at once conceptually rigorous and spiritually upbuilding.
- *Merold Westphal*

$49.95 | 6 x 9, 352 pages | Cloth | ISBN 1-932792-35-X

1-800-537-5487 B **baylorpress.com**

...on the path of faith and understanding

THINK

Wilfrid Sellars
Willem deVries

Philosophy Now
0-7735-3051-7 $22.95 paper
0-7735-3050-9 $75.00 cloth
352pp

"A first-rate work on one of the central philosophical figures of the twentieth century. This book is a major achievement."
Paul Coates, University of Hertfordshire

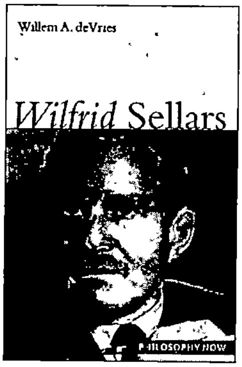
Central Works of Philosophy
Edited by John Shand

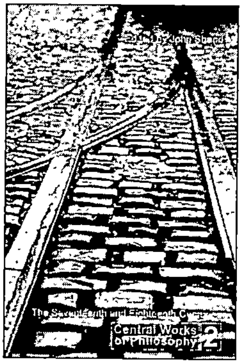
Volume 3: The Nineteenth Century
0-7735-3053-3 $24.95 paper
0-7735-3052-5 $85.00 cloth
256pp

Volume 4: The Twentieth Century: Moore to Popper
0-7735-3081-9 $24.95T paper
0-7735-3080-0 $85.00S cloth
320pp

A primer on the core texts of the Western philosophical canon aimed at undergraduate students and general readers.

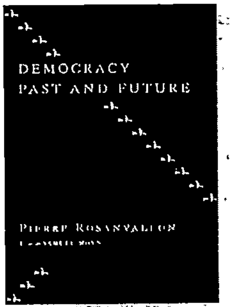
McGILL
QUEEN'S
UNIVERSITY
PRESS
www.mqup.ca



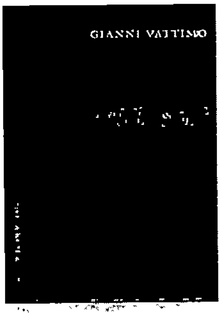


COLUMBIA, *Refer to us*
columbia.edu/cu/cup

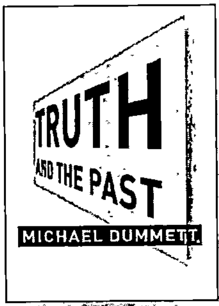
Dream I Tell You
Hélène Cixous
Translated by Beverley Bie Brahic
"[An] indefatigable and unique translation of the infinite world ... of the nocturnal dream."
— Jacques Derrida, *Geneses, Genealogies, Genres, and Genius*
cloth - $29.50



Democracy Past and Future
Pierre Rosanvallon
Edited by Samuel Moyn
"This collection ... will enrich our understanding of the perils and possibilities of democratic politics."
— Michael Sandel, *Harvard University*
cloth - $40.00



Dialogue with Nietzsche
Gianni Vattimo
Translated by William McCuaig
"This fine translation ... refreshes readers' appreciation of Vattimo's consistently insightful contributions to Nietzsche scholarship."
— Babette E. Babich, author of *Nietzsche's Philosophy of Science: Reflecting Science on the Ground of Art and Life*
cloth - $29.50



NOW IN PAPER
Truth and the Past
Michael Dummett
Foreword by Akeel Bilgrami
"This is the work of a great philosopher at the peak of his form."
— Hilary Putnam, *Harvard University*
paper - $19.00

CAMBRIDGE

NEW AND NOTEWORTHY FROM CAMBRIDGE

Economic Analysis, Moral Philosophy and Public Policy
Second Edition
Daniel M. Hausman and
Michael S. McPherson
$70.00: Hardback: 0-521-84629-3: 352pp
$28.99: Paperback: 0-521-60866-X

Kierkegaard and Socrates
A Study in Philosophy and Faith
Jacob Howland
$80.00: Hardback: 0-521-86203-5: 248pp

Ayn Rand's Normative Ethics
The Virtuous Egoist
Tara Smith
$80.00: Hardback: 0-521-86050-4: 328pp

The Cognitive Structure of Scientific Revolutions
Hanne Anderson, Peter Barker, and
Xiang Chen
$70.00: Hardback: 0-521-85575-6: 216pp

Sextus Empiricus: Against the Logicians
Richard Bett
Cambridge Texts in the History of Philosophy
$70.00: Hardback: 0-521-82497-4: 254pp
$27.99: Paperback: 0-521-53195-0

Extending Mechanics to Minds
The Mechanical Foundations of Psychology and Economics
Jon Doyle
$95.00: Hardback: 0-521-86197-7: 480pp

Best Book in 2005 on Social Philosophy by The North American Society for Social Philosophy

Crimes Against Humanity
A Normative Account
Larry May
Cambridge Studies in Philosophy and Law
$75.00: Hardback: 0-521-84079-1: 326pp
$27.99: Paperback: 0-521-60051-0

Named 2006 Book of Distinction by the Templeton Foundation

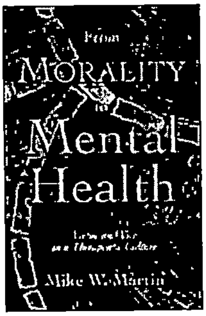
Bodies and Souls, or Spirited Bodies?
Nancey Murphy
Current Issues in Theology
$65.00: Hardback: 0-521-85944-1: 164pp
$22.99: Paperback: 0-521-67676-2

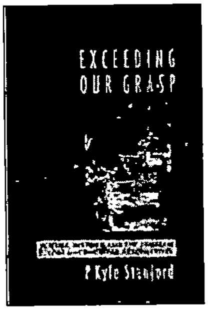
Prices subject to change

www.cambridge.org/us/philosophy/

CAMBRIDGE
UNIVERSITY PRESS

New from
Oxford





Becoming a Subject
Reflections in Philosophy and Psychoanalysis
MARCIA CAVELL

Marcia Cavell draws on philosophy, psychoanalysis, and the sciences of the mind in a fascinating and original investigation of human subjectivity. Cavell brings together for the first time reflections in philosophy, findings in neuroscience, studies in infant development, psychoanalytic theory, and clinical vignettes from her own psychoanalytic practice.

2006 $45 00

A Philosophy of Gardens
DAVID E. COOPER

Given the enthusiasm for gardens in human civilization ancient and modern, Eastern and Western, it is surprising that the subject has been so long neglected by modern philosophy. Now at last there is a philosophy of gardens. David Cooper identifies garden appreciation as a special human phenomenon distinct from both from the appreciation of art and the appreciation of nature.

2006 $29 95

From Morality to Mental Health
Virtue and Vice in a Therapeutic Culture
MIKE W. MARTIN

Morality and mental health are now inseparably linked in our view of character. In this wide-ranging, accessible book, Mike W. Martin asks: are we replacing morality with therapy, in potentially confused and dangerous ways, or are we creatively integrating morality and mental health?

(Practical and Professional Ethics)

2006 $29.95

Exceeding Our Grasp
Science, History, and the Problem of Unconceived Alternatives
P. KYLE STANFORD

The incredible achievements of modern scientific theories lead most of us to embrace scientific realism: the view that our best theories offer us at least roughly accurate descriptions of otherwise inaccessible parts of the world like genes, atoms, and the big bang. In *Exceeding Our Grasp,* Stanford argues that careful attention to the history of scientific investigation invites a challenge to this view that is not well represented in contemporary debates about the nature of the scientific enterprise.

2006 $45 00

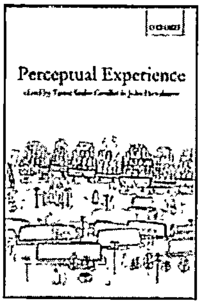
Epistemology Futures
Edited by STEPHEN HETHERINGTON

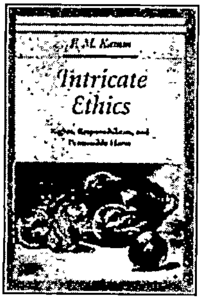
How might epistemology build upon its past and present, so as to be better in the future? *Epistemology Futures* takes bold steps towards answering that question.

2006 paper $35 00 cloth $99.00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS





New from
Oxford

Perceptual Experience

Edited by TAMAR SZABO GENDLER
and JOHN HAWTHORNE

Perceptual Experience is comprised of papers written specially for this volume, and they cover a broad range of topics dealing with sensation and representation, consciousness and awareness, and the connections between perception and knowledge and between perception and action. Contributors include John Campbell, David J. Chalmers, Tim Crane, Fred Dretske, Tamar Szabo Gendler, Anil Gupta, John Hawthorne, Susan Hurley, Mark Johnston, Geoffrey Lee, Eric Lormand, M. G. F. Martin, Alva Noë, Jesse J. Prinz, Sydney Shoemaker, Susanna Siegel, and Michael Tye.

February 2006 paper $35.00 cloth $99.00

Intricate Ethics

Rights, Responsibilities, and Permissible Harm

F. M. KAMM

In *Intricate Ethics*, Kamm questions the moral importance of some non-consequentialist distinctions and then introduces and argues for the moral importance of other distinctions. The first section provides a general introduction to non-consequentialist ethical theory; the second deals with the notions of moral status and rights; the third takes up the notions of responsibility and complicity; and the fourth section analyzes the views of others in the non-consequentialist and consequentialist camps.

March 2006 $65.00

Analyzing Oppression

ANN E. CUDD

Analyzing Oppression presents a new, integrated theory of social oppression, which tackles the fundamental question that no theory of oppression has satisfactorily answered: if there is no natural hierarchy among humans, why are some cases of oppression so persistent? Cudd argues that the explanation lies in the coercive co-opting of the oppressed to join in their own oppression.

March 2006 paper $35.00 cloth $74 00

My Way

Essays on Moral Responsibility

JOHN MARTIN FISCHER

This is a selection of essays on moral responsibility that represent the major components of John Martin Fischer's overall approach to freedom of the will and moral responsibility.

February 2006 $45.00

Choosing Children

The Ethical Dilemmas of Genetic Intervention

JONATHAN GLOVER

Genetic and reproductive technology now offers us the possibility of choosing what kinds of children we do and don't have. Should we welcome this power, or should we fear its implications? The renowned moral philosopher and best-selling author Jonathan Glover suggests how we should try to deal with this urgent problem.

March 2006 $15 95

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS

COMING SOON:

THE JOURNAL OF PHILOSOPHY

ONLINE

The Journal of Philosophy will be adding
a print and online subscription option for institutions
to begin Fall 2006

SUBSCRIPTION OPTIONS

Libraries and Institutions (print and online)	$180.00
Libraries and Institutions (print only)	$100.00
Individuals	$45.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

We expect to launch this new feature with our September 2006 issue.
Please consult our website at www.journalofphilosophy.org
for more details.

Batla
9/11/06
CAMPUS

THE JOURNAL OF PHILOSOPHY

VOLUME CIII, NUMBER 5
MAY 2006

page
215 *What Do We Want from a Theory of Justice?* Amartya Sen
239 *Has Game Theory Been Refuted?* Francesco Guala
264 NEW BOOKS: REPRINTS
265 NEW BOOKS: TRANSLATIONS

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Christopher Peacocke, Carol Rovane, Achille C. Varzi. *Consulting Editors*: David Albert, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2006

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2006 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIII, NO. 5, MAY 2006

WHAT DO WE WANT FROM A THEORY OF JUSTICE?*

I begin from the general Rawlsian position that the interpretation of justice is linked with public reasoning. The focus has to be, in John Rawls's words, on "a public framework of thought" that provides "an account of agreement in judgment among reasonable agents."[1] Rawls outlines this demand in terms of avoiding what he calls "a personal slant":

> We do not look at the social order from our situation but take up a point of view that everyone can adopt on an equal footing. In this sense we look at society and our place in it objectively: we share a common standpoint along with others and do not make our judgments from a personal slant.[2]

The bearing of *public reasoning* on the theory of justice leads to two further inquiries: *What is the relevant public?* and *On what questions should the reasoning concentrate?* The former query concerns the range of points of view that should count in public reasoning (for exam-

---

* For helpful discussion, I am grateful to Sabina Alkire, Kenneth Arrow, Akeel Bilgrami, Sissela Bok, Joshua Cohen, Peter Hammond, Isaac Levi, Thomas Nagel, Prasanta Pattanaik, Mozaffar Qizilbash, Ingrid Robeyns, Emma Rothschild, Carol Rovane, Maurice Salles, Thomas Scanlon, Patrick Suppes, Kotaro Suzumura, Philippe Van Parijs, and to the participants of seminars at Harvard University, Stanford University, and Yale University, where earlier versions of this paper were presented.

[1] See Rawls, *Political Liberalism* (New York: Columbia, 1993), pp. 110–13.

[2] Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971), pp. 516–17. This corresponds to page 453 of the revised edition of the book (Cambridge: Harvard, 1999). On the relation between political objectivity and public reasoning, see also Bruce A. Ackerman, *Social Justice in the Liberal State* (New Haven: Yale, 1980); Joshua Cohen, "An Epistemic Conception of Democracy," *Ethics*, XCVII (1986–87): 26–38; Seyla Benhabib, ed., *Democracy and Difference: Contesting the Boundaries of the Political* (Princeton: University Press, 1996); Amy Gutmann and Dennis Thompson, *Democracy and Disagreement* (Cambridge: Harvard, 1996); Thomas Scanlon, *What We Owe to Each Other* (Cambridge: Harvard, 1998).

0022-362X/06/0305/215–38

ple, whether they must all come from inside a given political state), while the latter relates to the subject matter of public reasoning, in particular what are the questions to be answered for a satisfactory theory of justice? The two issues, I will argue, are linked, and together they lead us to the foundational question: *What do we want from a theory of justice?*

I have begun by drawing on Rawls's lead on the basic connection between objectivity, public reasoning, and the theory of justice. However, I have to argue for a rather different way of pursuing that connection, departing not only from the substantive content of the Rawlsian theory of justice but also from Rawls's diagnosis of the very requirements of a theory of justice, including the subject matter of public reasoning and the reach and coverage of public participation.

THE TRANSCENDENTAL VERSUS THE COMPARATIVE

I begin with the issue of the subject matter of a satisfactory theory of justice. In his analysis of "justice as fairness," Rawls takes the principal question to be: *What is a just society?* Indeed, in most theories of justice in contemporary political philosophy, that question is taken to be central. This leads to what can be called a "transcendental" approach to justice, focusing—as it does—on identifying perfectly just societal arrangements. In contrast, what can be called a "comparative" approach would concentrate instead on ranking alternative societal arrangements (whether some arrangement is "less just" or "more just" than another), rather than focusing exclusively—or at all—on the identification of a fully just society. The transcendental and comparative approaches are quite distinct, and as will be presently discussed, neither approach, in general, subsumes or entails the other.[3]

The transcendental approach to justice is not new (it can be traced at least to Thomas Hobbes), but recent contributions have done much to consolidate the reliance on this approach. In his investigation of "justice as fairness," Rawls explores in depth the nature of an entirely just society seen in the perspective of contractarian fairness. Rawls's investigation begins with identifying the demands of fairness through exploring an imagined "original position" in which the members of the society are ignorant of their respective individual characteristics including their own comprehensive preferences. The principles of justice that emerge in the original position are taken to be impartial because they are chosen by the persons involved under a "veil of ignorance," without knowledge of their individual identities in the society with specific vested interests and particular priorities.

---

[3] Social choice theory, pioneered in the modern form by Kenneth Arrow (*Social Choice and Individual Values* (New York: Wiley, 1951)), is based on a fundamentally comparative approach. The analytical priority of the comparative can also be seen in the early formulations of social choice in the eighteenth century, particularly in Marquis de Condorcet, *Essai sur l'Application de l'Analyse à la Probabilité des Decisions rendues à la Pluralité des Voix* (Paris: L'Imprimerie Royale, 1785).

Later on in this paper, I shall discuss some limitations of this understanding of the demands of fairness (and ask whether the points of view to be considered must all come from the population of a given state), but the immediate point to note in the context of understanding the transcendental approach is that the fairness exercise is aimed entirely at identifying appropriate principles for a fully just society and at isolating the institutional needs for the basic structure of such a society. The working of these institutions, in turn, leads to further societal decisions at later stages in the Rawlsian system, for example through appropriate legislation (in what Rawls calls "the legislative stage"). The sequence moves forward step by step on firmly specified lines, with elaborately characterized unfolding of completely just societal arrangements.

Despite the standing and widespread use of the transcendental approach, the intellectual interest in, and practical relevance of, comparative questions about justice are hard to deny. Investigation of different ways of advancing justice in a society (or in the world), or of reducing manifest injustices that may exist, demands comparative judgments about justice, for which the identification of fully just social arrangements is neither necessary nor sufficient. To illustrate the contrast involved, it may well turn out that in a comparative perspective, the introduction of social policies that abolish slavery, or eliminate widespread hunger, or remove rampant illiteracy, can be shown to yield an advancement of justice. But the implementation of such policies could still leave the societies involved far away from the transcendental requirements of a fully just society (since transcendence would have other demands regarding equal liberties, distributional equity, and so on).

The grand partition between the "just" and the "nonjust," which is what a theory of transcendental justice yields, would leave the society on the "nonjust" side even after the reform, despite what can be seen, in a comparative perspective, as a justice-enhancing change. Some nontranscendental articulation is clearly needed. To take another type of example, instituting a system of public health insurance in the United States that does not leave tens of millions of Americans without any guarantee of medical attention at all may be judged to be an advancement of justice, but such an institutional change would not turn the United States into a "just society" (since there would remain a hundred other transgressions still to remedy).

A transcendental approach cannot, on its own, address questions about advancing justice and compare alternative proposals for having a more just society, short of proposing a radical jump to a perfectly just world. Indeed, the answers that a transcendental approach to justice gives—or can give—are quite distinct and distant from the type of concerns that engage people in discussions on justice and injustice in the world, for example, iniquities of hunger, illiteracy, torture, arbitrary incarceration, or medical exclusion as particular social features that need remedying. The focus of these engagements tends to be on the ways and means of advancing justice—or reducing injustice—in the world by remedying these inequities, rather than on looking only for the simultaneous fulfilment of the entire cluster of perfectly just societal arrangements demanded by a particular transcendental theory.

POSSIBLE DEFENSE OF A TRANSCENDENTAL APPROACH

The argument so far has been, in an important sense, too easy. Surely transcendental answers cannot be all we want from a theory of justice. But there might well be—this is a matter to be investigated—some less obvious connection, some relationship between the transcendental and the comparative that could make the transcendental approach the right way of proceeding to comparative assessments. The formal remoteness of the transcendental approach from the invoking of the idea of justice in debates and discussions on practical affairs does not in itself indicate that the transcendental approach cannot be the right approach.

Thus, at least two further questions must be addressed, related to the possibility, respectively, of (1) the sufficiency, and (2) the necessity, of the transcendental approach for making comparative judgments about justice. First, can the answers to transcendental queries take us indirectly to comparative assessments of justice as well, in particular through comparisons of "distances" from transcendence at which particular sets of societal arrangements respectively stand? Second, can it be the case that the transcendental question ("What is a just society?") has to be answered first, as an essential requirement, for a cogent and well-founded theory of comparative justice, which would otherwise be foundationally disjunctive and frail? For an adequate critique of the transcendental approach to justice from the comparative perspective, we have to assess these possibilities critically.

Implicit beliefs in the sufficiency or the necessity (or both) of a transcendental approach for comparative assessment clearly have had a powerful role in the widespread belief that the transcendental approach is crucial for the entire theory of justice. Indeed, even in social choice theory, where the analytical framework is firmly relational and altogether grounded on pairwise comparisons, the investigations of justice in particular has been standardly elongated to move relentlessly from the basic comparative rankings to the identification of transcendental justice (often in the Rawlsian mold).[4] In arguing for a more robustly comparative approach to justice, with which this paper is concerned (and for which social choice theory can play, I would suggest, an important role), it would be necessary to examine whether comparative conclusions either *follow from*, or *need*, some transcendental identification.

DOES TRANSCENDENTAL SPECIFICATION YIELD COMPARATIVE RANKINGS?

I begin with the issue of *sufficiency*. Does a transcendental approach produce, as a by-product, relational conclusions that are ready to be drawn out, so that transcendence may end up giving us a great deal more than its overt form articulates? In particular, is the specification of an entirely just society sufficient to give us rankings of departures from justness in terms of comparative "distances" from perfection, so that a transcendental identification might immediately entail comparative gradings as well?

The answer here is a firm no. The main difficulty lies in the fact that there are different features involved in identifying distance, related, among other distinctions, to (1) different fields of departure, (2) varying dimensionalities of transgressions within the same general field, and (3) diverse ways of weighing separate infractions. The identification of transcendence does not yield any means of addressing these problems to arrive at a relational ranking of departures from transcendence.

For example, in the context of the Rawlsian analysis of the just society, departures may occur in many different spaces. They can include the breaching of liberty, which, furthermore, can involve diverse violations of distinctive liberties (many of which figure in Rawls's capacious coverage of liberty and its priority). There can also be violations—again in possibly disparate forms—of the demands of equity in the distribution of primary goods (there can be many different departures from the demands of the Difference Principle which forms a part of Rawls's second principle). Similarly, diverse transgressions can occur in other transcendental theories of justice (for example, those that would replace the Rawlsian focus on "primary goods" in the Difference Principle by concentrating respectively on "capabilities" or "resources" or "opportunities," or some other way of formulating the allocational and distributional needs of transcendental justice).

---

[4] The comparative conclusions have typically been pieced together to arrive at some ultimately transcendental claims (a translation that is possible with rather restrictive and limiting assumptions, as will be discussed presently). See, for example, Sen, *Collective Choice and Social Welfare* (San Francisco: Holden-Day, 1970); Kotaro Suzumura, *Rational Choice, Collective Decisions, and Social Welfare* (New York: Cambridge, 1983); Peter J. Hammond, "Equity, Arrow's Conditions and Rawls' Difference Principle," *Econometrica*, XLIV (1976): 793–804; Claude d'Aspremont and Louis Gevers, "Equity and the Informational Basis of Collective Choice," *Review of Economic Studies*, XLVI (1977): 199–209; Arrow, "Extended Sympathy and the Possibility of Social Choice," *American Economic Review*, LXVII (1977): 219–25. The influence of the Rawlsian perspective on justice has been very strong on the application of social choice theory, despite the basic distance between the respective comparative and transcendental approaches.

There are also disparate ways of assessing the extent of each such discrepancy and of appraising the comparative remoteness of actual distributions from what the principles of full justice would demand. Further, we have to consider departures in procedural equity (such as infringements of fair equality of public opportunities or facilities), which figure within the domain of Rawlsian demands of justice (in the first part of second principle). To weigh these procedural departures *against* infelicities of emergent patterns of interpersonal distribution (for example, distributions of primary goods), which also figure in the Rawlsian system, would require distinct specification—possibly in axiomatic terms—of relative importance or significance (or "trade-offs" as they are sometimes called in the crude vocabulary of multidimensional assessment). But these extensions, helpful as they would be, lie well beyond the specific exercise of the identification of transcendence, and are indeed the basic ingredients of a "comparative" rather than a "transcendental" approach to justice. The characterization of spotless justice does not entail any delineation whatever of how diverse departures from spotlessness can be compared and ranked.

The absence of such comparative implications is not, of course, an embarrassment for a transcendental theory of justice, seen as a freestanding achievement. The relational silence is not, in any sense, an internal difficulty of a transcendental theory of justice. Indeed, some pure transcendentalists would be utterly opposed even to flirting with gradings and comparative assessments, and may quite plausibly shun relational conclusions altogether. They may point in particular to their understanding that a "right" social arrangement must not, in any way, be understood as a "best" social arrangement, which could open the door to what is sometimes seen as the intellectually mushy world of graded evaluations in the form of "better" or "worse" (linked with the relationally superlative "best"). The absoluteness of the transcendental "right"—against the relativities of the "better" and the "best"—may well have a powerfully reasoned standing of its own. But it does not help at all in *comparative assessments* of justice.

To be sure, members of any polity can contemplate how a gigantic and totally comprehensive reorganization may be brought about, moving us *at one go* to the ideal of a fully just society. A no-nonsense transcendental theory can serve, in this sense, as something like the "grand revolutionary's complete handbook." But that handbook would not be much invoked in the debates on justice in which we are constantly engaged, which focus on how to reduce the manifold injustices that characterize the world.[5]

Even if we think of transcendence not in the "gradingless" terms of "right" social arrangements, but in the graded terms of the "best" social arrangements, the identification of the best does not, in itself, tell us much about the full grading, such as how to compare two nonbest alternatives. The identification of the best does not specify a unique ranking with respect to which the best stands at the pinnacle; indeed the same best may go with a great many different rankings with the same pinnacle. To consider an analogy, the fact that a person regards the *Mona Lisa* as the best picture in the world, does not reveal how she would rank a Gauguin against a Van Gogh. The search for transcendental justice is an engaging exercise in itself, but irrespective of whether we think of transcendence in terms of the gradeless "right" or in the framework of the graded "best," it does not tell us much about the comparative merits of many—indeed typically most—of the different societal arrangements.

IS A TRANSCENDENTAL THEORY NECESSARY FOR COMPARISONS OF JUSTICE?

I now take up the second question, concerning the hypothesis that the identification of the best is necessary, even if not sufficient, to rank any two alternatives in terms of justice. In the usual sense of necessity, this would be a somewhat odd possibility. In the discipline of comparative judgments in any field, relative assessment of two alternatives tends in general to be a matter between them, without there being the necessity of beseeching the help of a third—

---

[5] It is worth noting here that the diagnosis of injustice does not demand a unique identification of the "just society," since many different identifications of perfectly just social arrangements may all agree on the diagnosis of a remediable deficiency of a particular social arrangement (say, with manifest hunger or illiteracy or medical neglect).

"irrelevant"—alternative. Indeed, it is not at all obvious why in making the judgment that some social arrangement x is better than an alternative arrangement y, we have to invoke the identification that some quite different alternative z is the "best" or the "right" social arrangement. In arguing for a Picasso over a Dali we do not need to get steamed up about identifying the perfect picture in the world, which would beat the Picassos and the Dalis and all other paintings in the world.

It might, however, be thought that the analogy with aesthetics is problematic since a person might not even have any idea of a perfect picture, in a way that the idea of a perfectly just society has appeared to be identifiable, in transcendental theories of justice. I will argue later on that the existence of a best, or a transcendent, alternative is actually not guaranteed even in the field of justice, but I am ready to proceed, for the moment, on the presumption that such an identification can somehow be made. However, despite this tentative acceptance, the existence of an identifiably inviolate, or best, alternative does not indicate that it is necessary (or indeed useful) to refer to it in judging the relative merits of two other alternatives. For example, we may indeed be willing to accept, with great certainty, that Everest is the tallest mountain in the world, completely unbeatable in terms of stature by any other peak, but that understanding is neither needed, nor particularly helpful, in comparing the heights of, say, Kanchenjunga and Mont Blanc. There would be something very deeply odd in a general belief that a comparison of any two alternatives cannot be sensibly made without a prior identification of a supreme alternative.

Thus, the hypothesis of necessity in the standard sense would be hard to sustain. There is, however, a weaker form of the hypothesis of necessity, which merely asserts that if comparative assessments can be systematically made, then that discipline must also be able to identify the very best. The claim, in this case, would be not so much that two alternatives cannot be compared in terms of justice without first knowing what the best or the perfect alternative is, but that the comparative ranking of the different alternatives must *inter alia* also be able to identify the answer to the transcendental question regarding the perfectly just society. Or, to put it in another way, if the transcendental question cannot be answered, then nor can be the comparative. This understanding of necessity would not vindicate the need to go *via* the transcendental approach to comparative assessments, but it would at least give transcendental identification a necessary presence in the theory of justice. We have to examine this considerably weaker claim of "necessity" as well.

COMPARATIVES WITHOUT TRANSCENDENCE

Would a sequence of pairwise comparisons invariably lead us to the very best? That presumption has some appeal, since the superlative might indeed appear to be the natural end point of a robust comparative. But this conclusion would, in general, be a *non sequitur*. In fact, it is only with a "well-ordered" ranking (for example, a complete and transitive ordering over a finite set) that we can be sure that the set of pairwise comparisons must also identify a "best" alternative.

We must, therefore, ask: How complete should the assessment be, for it to be a systematic discipline? In the "totalist" approach that characterizes the standard theories of justice (including Rawls's), incompleteness tends to appear as a failure, or at least as a sign of the unfinished nature of the exercise. Indeed, the survival of incompleteness is sometimes seen as a defect of a theory of justice, which calls into question the positive assertions that such a theory makes. In fact, however, a theory of justice that makes systematic room for incompleteness allows one to arrive at possibly quite strong judgments (for example, about the injustice of continuing famines in a world of prosperity, or of persistently grotesque subjugation of women), without having to find highly differentiated assessment of every political and social arrangement in comparison with every other arrangement (for example, addressing such questions as: Is a top income tax rate of 45 percent more just or less just than a top rate of 46 percent?)

I have discussed elsewhere why a systematic and disciplined theory of normative evaluation, including assessment of social justice, need not take a "totalist" form.[6] Incompleteness may be of the lasting kind for several different reasons, including unbridgeable gaps in information, and judgmental unresolvability involving disparate considerations that cannot be entirely eliminated, even with full information. For example, it may be hard to resolve the overall balance of the comparative claims of equity considerations that lie behind Rawlsian lexicographic maximin, compared with, say, sum-ranking in a gross or equity-adjusted form.[7] And yet, despite such durable incompleteness, we may still be able to agree readily that there is a clear social injustice involved in the persistence of endemic hunger or exclusion from medical access, which calls for a well-specified remedying for the *advancement* of justice (or reduction of injustice), even after taking note of the costs involved. Similarly, we may acknowledge the possibility that liberties of different persons may, to some extent, conflict with each other (so that any fine-tuning of the demands of equal liberty may be hard to work out), and yet strongly agree that torturing accused people would be an unjust violation of liberty and that this injustice calls for an urgent rectification.

---

[6] Sen, "Consequential Evaluation and Practical Reason," this JOURNAL, XCVII, 9 (September 2000): 477–572. See also Sen, *Collective Choice and Social Welfare*; "Maximization and the Act of Choice," *Econometrica*, LXV (1997): 745–79; and "The Possibility of Social Choice," *American Economic Review*, LXXXIX (1999): 349–78; and also "Incompleteness and Reasoned Choice," *Synthese*, CXL (2004): 43–59. See also Isaac Levi's response to the last, in "Amartya Sen," in the same number of *Synthese*, pp. 61–67.

[7] The vast literature on this include, among other contributions, S.-Ch. Kolm, "The Optimum Production of Social Justice," in J. Margolis and H. Guitton, eds., *Public Economics* (London: Macmillan, 1969), pp. 145–200; A.B. Atkinson, "On the Measurement of Inequality," *Journal of Economic Theory*, II (1970): 244–63; James Mirrlees, "An

There is a further consideration that may work powerfully in the direction of making political room for incompleteness of judgments about social justice, even if it were the case that every person had a complete ordering over the possible social arrangements. Since a theory of justice invokes agreement between different parties (for example, in the "original position" in the Rawlsian framework), incompleteness can also arise from the possibility that different persons may continue to have some differences (consistently with agreeing on a lot of the comparative judgments). Even after vested interests and personal priorities have been somehow "taken out" of consideration through such devices as the "veil of ignorance," there may remain possibly conflicting views on social priorities, for example in weighing the claims of need over entitlement to the fruits of one's labour.

Conflicts of distributive principles that are hard to eradicate can be illustrated with an example, which I have discussed in another context. The example is concerned with the problem of deciding which of three children should get a flute about which they are quarrelling.

Exploration of the Theory of Optimal Income Taxation," *Review of Economic Studies*, XXXVIII (1971): 175–208; Sen, *On Economic Inequality* (New York: Oxford, 1973; enlarged edition with a new addendum jointly written with James Foster, 1997); Claude d'Aspremont and Louis Gevers, "Equity and the Informational Basis of Collective Choice," XLIV (1977): 199–209; Eric Maskin, "Decision-making under Ignorance with Implications for Social Choice," *Theory and Decision*, XI (1979): 319–37; Kevin W.S. Roberts, "Interpersonal Comparability and Social Choice Theory," *Review of Economic Studies*, XLVII (1980): 421–39; Charles Blackorby, David Donaldson, and John Weymark, "Social Choice with Interpersonal Utility Comparisons: A Diagrammatic Introduction," *International Economic Review*, XXV (1984): 327–56; d'Aspremont, "Axioms for Social Welfare Ordering," in Leonid Hurwicz, David Schmeidler, and Hugo Sonnenschein, eds., *Social Goals and Social Organization: Essays in Memory of Elisha Pazner* (New York: Cambridge, 1985), pp. 19–76; and d'Aspremont and Gevers, "Social Welfare Functionals and Interpersonal Comparability," Blackorby, Walter Bossert, and Donaldson, "Utilitarianism and the Theory of Justice," and Bhaskar Dutta, "Inequality, Poverty and Welfare," the last three in Arrow, Sen, and Suzumura, eds., *Handbook of Social Choice and Welfare*, Volume 1 (Amsterdam: North-Holland, 2002), chapters 10–12.

Child *A* is the only one of the three who knows how to play the flute (the others do not deny this); child *B* is the only one without any toys of his own (the other two concede that they are much richer and well supplied with engaging amenities); child *C* has worked hard to make the flute all on his own (the others confirm this). Theorists of different persuasions—utilitarian or egalitarian or libertarian—may believe that a just resolution can be readily spotted here, though, alas, they would respectively see totally different resolutions as being exactly right. The main point to note in the present context is that the different resolutions all have serious arguments in support of them, and we may not be able to identify exactly one of the alternative arguments as being the only one (to invoke Thomas Scanlon's criterion) that "could be justified to others on grounds that they, if appropriately motivated, could not reasonably reject."[8]

Even when each of the parties involved has his or her own complete specification of justice, the "intersection" between the rankings—that is the *shared* beliefs of the different parties—can yield a partial ranking, if the judgments are not all congruent.[9] The acceptability of evaluative incompleteness is indeed a central subject in social choice in general, and it is relevant to theories of justice as well, even though Rawlsian and other theories assert (and it *is* an assertion rather than something that is established in any clear way) that a full agreement will definitely emerge in the "original position" and in other such formats.[10]

Indeed, for reasons both of incomplete individual evaluations and of incomplete congruence of individual assessments, incompleteness may be a hardy feature of judgments of social justice. This can be problematic for the identification of a perfectly just society, and make transcendental conclusions difficult to derive.[11] And yet, such incompleteness would not prevent making comparative judgments of justice in a great many cases, where there might be fair agreement on particular pairwise rankings, about how to enhance justice and reduce injustice. A partial ordering can be very useful without being able to lead to any transcendental identification of a fully just society.

---

[8] Scanlon, *What We Owe to Each Other*, p. 5; see also his "Contractualism and Utilitarianism," in Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (New York: Cambridge, 1982), pp. 103–28.

[9] The formal characteristics of "intersection partial orderings" are discussed in Sen, *On Economic Inequality*, and "Maximization and the Act of Choice."

[10] For an early expression of skepticism about the plausibility of unanimous judgments in the "original position," see my joint essay with W.G. Runciman, "Games, Justice and the General Will," *Mind*, LXXIV (September 1965): 554–62.

[11] On a mathematical point, it must be acknowledged that a transitive but incomplete ordering over a finite set will invariably yield one or more "maximal" elements, in the sense of there being no element that is better than a maximal element. Maximality does not, however, guarantee the existence of a best element. The foundational nature of the mathematical distinction involved and its significant implications are investigated in N. Bourbaki, *General Topology*, Parts I and II, English translation (Reading, MA:

The question "What is a just society?" is, therefore, not a good *starting point* for a useful theory of justice. To that has to be added the further conclusion that it may not be a plausible *end point* either. A systematic theory of comparative justice does not need, nor does it necessarily yield, an answer to the question "What is a just society?"

INSTITUTIONAL BARRIERS AND TRANSCENDENTAL SILENCE

I turn now to a different—though not unrelated—feature of the transcendental approach to justice, in particular the extremely demanding institutional requirements of accomplishing pristine justice. The achievement of a fully just society would require a plethora of institutions, including the unfettered operations of a sovereign state. Some of these institutions are absent or defective in many countries in the world; nor can these countries readily establish them. Even without the possibility of setting up some of these institutions, it is, of course, possible to advance justice—or to reduce injustice—to a considerable extent, but while that is good enough for applying the comparative approach to justice, it does not yield the achievement of transcendental justice. If such spotless justice were the only focus of attention in a theory of justice, then the institutional preconditions would form a kind of "entry barrier," leading to an abstinence from applying justice theory to situations in which those exacting institutional demands are not only not currently met but cannot be met in the foreseeable future.

The institutional preconditions would be particularly hard to meet in dealing with, say, problems of global justice. The claim that we need a sovereign state to apply the principles of justice—a claim that was well articulated by Thomas Hobbes—is substantially connected with the elaborate institutional demands of a transcendental understanding of justice. Thomas Nagel's strongly argued dismissal of the

---

Addison-Wesley, 1966), and *Theory of Sets* (Reading, MA: Addison-Wesley, 1968). The presence of maximal elements is sufficient for reasoned choice of an alternative that is no worse than any other. It will not, however, in general allow the unfolding of a perfectly just social arrangement, not to mention a unique arrangement of perfect justice (as in the Rawlsian and other transcendental systems). On the far-reaching relevance of the distinction between maximality and optimality (and transcendence), see Sen, "Internal Consistency of Choice," *Econometrica*, LXI (1993): 495–521, and "Maximization and the Act of Choice," *Econometrica*, LXV (1997): 745–79.

relevance of "the idea of global justice" draws on his understanding that these extensive institutional demands cannot be met at the global level at this time. As he puts it, "It seems to me very difficult to resist Hobbes's claim about the relation between justice and sovereignty," and "if Hobbes is right, the idea of global justice without a world government is a chimera."[12] In the global context, Nagel concentrates, therefore, on clarifying other demands, distinguishable from the demands of justice, such as "minimal humanitarian morality" (which "governs our relations to all other persons"); and also to long-run strategies for radical change in institutional possibilities ("I believe the most likely path toward some version of global justice is through the creation of patently unjust and illegitimate global structures of power that are tolerable to the interests of the most powerful current nation-states").[13]

In the Rawlsian approach too, the application of a theory of justice requires an extensive cluster of institutions that determines the basic structure of a fully just society. Not surprisingly, Rawls actually abandons his own principles of justice when it comes to the assessment of how to go about thinking about global justice. In a later contribution, *The Law of Peoples*, Rawls invokes a "second original position," with a fair negotiation involving representatives of different polities—or different "peoples" as Rawls call them—who serve as parties under this second veil of ignorance.[14] However, Rawls does not try to derive principles of justice that might emanate from this second original position, and concentrates instead on certain general principles of humanitarian behavior. The Rawlsian vehicle of justice that would take us rapidly forward in pursuit of some justice in a justiceless world remains stalled and stationary in the wintry morning of a world without a global state.

To be sure, Rawls need not agree that the world is really unjust if he remains fully attached to the belief that the concept of justice does not apply at the global level. Nagel, on the other hand, seems definitely convinced that "we do not live in a just world" (the opening sentence of "The Problem of Global Justice" (*op. cit.*, p. 113)). While I firmly agree with that conclusion (given what I see as the role of a theory of justice, the subject matter of this paper), it is not entirely clear to me how Nagel can make statements of this kind given his conviction that the idea of global justice is "a chimera."[15]

---

[12] Nagel, "The Problem of Global Justice," *Philosophy and Public Affairs*, XXXIII (2005): 113–47, here p. 115.

[13] Nagel, "The Problem of Global Justice," pp. 130–33, 146–47. An important critique of what Joshua Cohen and Charles Sabel describe as Nagel's "statism" can be found in their "Extra Rempublicam Nulla Justitia?," *Philosophy and Public Affairs*, XXXIV (Spring 2006): 147–75, in which see also A.L. Julius, "Nagel's Atlas," pp. 176–92.

[14] Rawls, *The Law of Peoples* (Cambridge: Harvard, 1999).

The challenge of making the world less unjust, thus, remains unaddressed within the transcendental approach. It is not, however, at all clear why we should be reduced to silence, so far as justice is concerned, merely because the reach of institutional possibilities does not prepare us for transcendental justice. The question how global justice can nevertheless be advanced remains pertinent to ask, unless we are forcibly removed from the territory of justice on the ground that transcendental justice is the only satisfactory—or the only understandable—idea of justice. Indeed, that question about advancement of justice (or about reduction of global injustice) can be a fruitful part of the subject matter of the deliberative framework of public reasoning. A "public framework of thought," which Rawls has taught us to value and use, should not become wholly inoperative merely because the institutional demands of a perfectly just society have turned out to be infeasible.

There is, thus, a real tension between making good use of public reasoning (drawing on Rawls's powerful, general arguments in that direction) and remaining silent whenever some exacting institutional conditions needed for transcendental justice cannot be entirely fulfilled (a conclusion that emerges from the special transcendental form that Rawls gives to his theory of justice). There is a serious loss here. Public deliberation can be important and useful both in pursuing institutional reform (even when the totality of the institutional reforms needed for perfection cannot be carried out) and in examining what can be done to reduce injustice—a basic question in the comparative approach to justice—even when not all the right institutions are in place.[16] A non-Rawlsian comparative (rather than transcendental) approach to justice can be a good conceptual base for such—essentially Rawlsian—public reasoning.

TRANSCENDENTAL LOCALISM AND GLOBAL CONCERNS

The importance of public reasoning for dealing with global problems of justice is a subject of importance on its own. However, it must be acknowledged that even though Rawls was a visionary leader of thought on the importance of public reasoning, he had considerable skepticism about the use of public reasoning at the global level. It is important to separate out two possible grounds for Rawls's reluctance. One issue, which has already been mentioned, is the inapplicability of the exacting framework of transcendental justice at the global level, because of institutional limitations. A second reason for Rawls's reluctance is his insistence on linking public reasoning with the contractarian format of the "original position." This involves a devised deliberative exercise that would appear to be hard to apply beyond the limits of a particular society (or a particular "people," as Rawls defines this collectivity in his later works).[17]

---

[15] If Nagel is relying on a possible formal interpretation (by stretching a point) that in the absence of an applicable theory of justice it can indeed be said that "we do not live in a just world," then he could have, with equal ease, also said, "we do not live in an unjust world." That may not, however, have served quite so well as the opening sentence of Nagel's finely argued and sensitive essay, and there is, I would argue, reason to ask ourselves how this asymmetry arises if the idea of global justice is really such a chimera.

[16] On this see my "Elements of a Theory of Human Rights," *Philosophy and Public Affairs*, XXXIV (Fall 2004): 315–56.

Rawls's statement about the need for a "common standpoint" which was quoted (from *A Theory of Justice*) at the beginning of this paper was immediately followed by the invoking of this particular conceptual device:

> Thus our moral principles and convictions are objective to the extent that they have been arrived at and tested by assuming this general standpoint and by assessing the arguments for them by the restrictions *expressed by the conception of the original position.*[18]

The deliberation thus takes the form of fair negotiation, in which the fairness of the reasoned negotiation is grounded on the demand that the reasoning occurs under a specially conceived veil of ignorance. But the participants in the deliberation are exactly the parties to the social contract for the society in question. A person's voice counts because he or she is directly involved in the social contract, which will "regulate the institutions" of the society of which he is a member.[19]

In contrast with this negotiational justification of the confinement to local points of view, there is a different approach to impartiality which brings in different voices, possibly even from "a distance" (to use Adam Smith's articulate phrase), precisely because these voices illuminate public decisions and help to make them impartial. In the terminology of conflict resolution, this is more like arbitration, rather than negotiation; the arbitrators need not themselves be parties to the dispute. I have argued elsewhere that the interpretation of fairness and impartiality through an understanding of "fair arbitration" is a serious rival to the route of "fair negotiation," which is the exclusive direction in which the contractarian feature of Rawls's transcendental approach—"justice as fairness"—proceeds.[20]

---

[17] It is, however, important to note that in the version of "contractualism" that Scanlon has developed and explored, which requires judgment about what it "would be reasonable for those affected by a principle to reject," citizenal confinement is neither necessary nor indeed sensible, since the policies and institutions of a nation, or state, or "people," as Rawls calls this collectivity, can significantly affect other people elsewhere. See Scanlon's *What We Owe to Each Other*, and also his earlier contribution, "Contractualism and Utilitarianism."

[18] Rawls, *A Theory of Justice* (1971), pp. 517, and (1999), p. 453; italics added.

[19] Rawls, *A Theory of Justice* (1971), p. 23.

The approach of fair arbitration is well exemplified by Smith's invoking of the perspectives of "impartial spectators." The impartial spectators are imagined observers who need not be members of the society, and their impartiality does not come, as in the Rawlsian system, exclusively—or even primarily—through the thought experiment of a veil of ignorance about the personal circumstances of individual members of a given society. Rather, the thought experiment by members of society, in the Smithian system of fair arbitration, invokes the judgments of disinterested observers who are not themselves parties to the societal decisions that are to be taken.

In itself this may not seem like a big difference, since both are merely thought experiments that must be undertaken, within the respective formats, by the people in the actual society. Also, there is nothing to prevent the imagined fair arbitrators from undertaking the exercise of placing themselves in the position of the parties involved under a devised veil of ignorance, so that fair arbitration can make good use of the insights that may come from fair negotiation.

However, there are two sources of substantial difference between the Smithian and the Rawlsian procedures. First, the contractarian approach standardly proceeds toward identifying the demands of transcendence (the principal inquiry in the original position is aimed at the demands of a just society), whereas Smith's impartial spectators are typically invoked for contrasting alternatives to throw light on specific issues of advancement or retardation of justice in a comparative approach. A second difference arises from the fact that the impartial observers may be imagined as coming from far as well as near, with questions being asked about how the decisional problem would look to those who may have had different social and institutional experience (a question of some importance, as Smith has argued). While the imagined impartial judges may find it useful to ask *inter alia* what things would look like had they actually been the involved parties (here the exercise would be, in effect, rather similar to the Rawlsian one), they could also be seen as bringing perspectives that are altogether different from the ones generally accepted in the local society and culture.

---

[20] Sen, "Open and Closed Impartiality," this JOURNAL, XCIX, 9 (September 2002): 445–69.

In the Rawlsian exercise, while there is a procedural requirement of ignorance of personal interests, personal aims, and personal circumstances, nothing is demanded about the knowledge and approbation of shared beliefs—and prejudices—of the society in which these individuals happen to live. In contrast, what the Smithian model of fair arbitration demands is that the people in any society must put in an effort to examine how their own practices and conventions would look to others, including people who are informed about, but not entirely reared in, that society.[21]

GLOBAL PERSPECTIVES, LOCAL INTERESTS, AND PAROCHIALISM

The reasons to go beyond the contracting parties in a nation state include at least three distinct concerns: (1) the interests of other people may be affected (for example, by national policies on "global warming," or for that matter what is called the "war against terror"); (2) the local parochialism shared by all (or nearly all) the persons within a given society may call for a distant challenge in the interest of objectivity; and (3) additional knowledge about what is feasible can be acquired from the experiences of other countries.[22]

On the first subject (that is the interests of people beyond borders which may be significantly affected), the arbitration approach allows the possibility of taking note of some of the broader concerns, including distributive ones, about global justice that have led in recent years to attempts to consider a "cosmopolitan" version of the original position, so that the interests of people in other countries, which may be influenced by policies in this country, are not neglected.[23] This is aimed at extending the reach of justice.

---

[21] The Rawlsian formulation of the "original position" suffers also from a further problem in specifying the allegedly fixed set of negotiators when the results of the negotiation can change the size and composition of the population involved. This problem of "inclusionary incoherence" (as I called it), which the Smithian approach does not have, was discussed, among other issues, in an earlier essay in this journal, "Open and Closed Impartiality." I shall not, however, further pursue here this rather different type of difficulty.

[22] On the last issue, it is worth mentioning that the understanding in different countries of social feasibilities (relevant, for example, for gender justice) and economic feasibilities (relevant, for example, for policies against poverty) is significantly influenced in the contemporary world by learning from the experiences—successes and failures—of other countries; on this see my *Development as Freedom* (New York: Knopf, 1999).

It is, however, important to see clearly that the invoking of impartial spectators from elsewhere does not make the Smithian exercise of arbitration by impartial spectators similar to the "cosmopolitan" version of the contractarian approach (which could take the form of including all the people in the world in one gigantic original position, yielding one huge global social contract for an entirely just world). The route of fair arbitration is fundamentally different from that of fair negotiation and of social contract, because of the way impartiality is interpreted.[24] The institutional demands of the contractarian procedure if aimed at the identification of transcendental justice take us immediately to the need for a sovereign nation, which was the bone of Thomas Nagel's contention. This is a problem that does not arise in a similar way in the case of invoking the device of the impartial spectators to assist in the assessment of justice in a comparative framework.[25]

Nevertheless, there are some similarities between the cosmopolitan version of the social contract and the Smithian invoking of distant observers. The Rawlsian exclusion of foreign nationals from having a voice in the assessment of policies of a country that have influences on the rest of the world is restrained in both the cosmopolitan social contract and in the Smithian exercise of invoking observers from far as well as near.

Second, the approach of the impartial spectator can bring in, *inter alia*, distant perspectives that are detached not only from the particular vested interests of individual citizens, but also from any parochialism of local beliefs that may be generally shared by all members of a given polity or community. One of the possible advantages of the route of fair arbitration is, thus, the greater versatility that the latter has, which can incorporate a systematic procedural challenge to the distortion of parochial convictions.

The avoidance of parochialism was, in fact, one of the principal reasons for Smith's insistence that the impartial spectators must *inter alia* represent perspectives from (as Smith put it) "a certain distance." Smith put the point thus: "we can do this in no other way than by endeavoring to view them with the eyes of other people, or as other people are likely to view them."[26]

---

[23] See, for example, Thomas Pogge, ed., *Global Justice* (Malden, MA: Blackwell, 2001). See also Deen Chatterjee, *The Ethics of Assistance: Morality and the Distant Needy* (New York: Cambridge, 2004).

[24] For example, it can be shown that the cosmopolitan version of the Rawlsian original position is also vulnerable to "inclusionary incoherence" in a way that the Smithian approach is not, as is discussed in my "Open and Closed Impartiality."

[25] Here again it must be noted that in the version of "contractualism" presented by Scanlon (*What We Owe to Each Other*), which has many affinities with the Smithian exercise of impartiality, the focus need not be on transcendence only.

In a chapter in *The Theory of Moral Sentiments* entitled "On the Influence of Custom and Fashion upon the Sentiments of Moral Approbation and Disapprobation," Smith argued that "the different situations of different ages and countries are apt...to give different characters to the generality of those who live in them, and their sentiments concerning the particular degree of each quality, that is either blamable or praise-worthy, vary, according to that degree which is usual in their own country, and in their own times."[27] One of Smith's illustrations of such parochial values was the tendency of all political commentators in ancient Greece, including sophisticated Athenians, to regard infanticide as perfectly acceptable social behavior. Even Plato and Aristotle did not depart from expressing approval, Smith noted, of this extraordinary practice which "uninterrupted custom had by this time...thoroughly authorized" in ancient Greece.

The Rawlsian device of losing information about personal identities in a given society, which does much to eliminate the influence of individual vested interests, does not provide any systemic way of avoiding prejudices that are broadly *shared* by everyone within a given society. As it happens, the localism of the contractarian approach is, in fact, reinforced by Rawls's insistence that the transcendental exercise in the original position should concentrate on "the basic structure" of "a closed society: that is, we are to regard it as self-contained and as having no relations with other societies."[28] Smith's argument that we must *inter alia* view our sentiments from "a certain distance from us" is motivated by the need to ask the question whether some appearance of justice is socially biased through the impact of entrenched tradition and local custom. Smith's actual example of infanticide remains distressingly relevant in some societies even today (though no longer in Greece), but there are also many other practices for which justice being seen to be done may usefully invoke, as Smith put it, "the eyes of the rest of mankind."[29]

While an American audience may find it easy to believe that distant perspectives may be usefully brought in, in the case of "backward" societies such as Sudan or Afghanistan, in which, for example, honor killings occur and adulterous women might be stoned to death, there may be no corresponding recognition of the need to do this for more advanced countries like the United States. However, well-established practices that receive widespread support within the borders might be seen as unacceptable to people in many other countries, spread across the world, from Europe to Japan. For example, plentiful use of capital punishments, with or without being accompanied by public jubilation, may need to be addressed not only by asking whether they appear "cruel and unusual" *within* the local U.S. culture, but also in "the eyes of the rest of mankind," which, Smith thought, must be invoked to understand whether "a punishment appears equitable" (*ibid.*, p. 104).

---

[26] Smith, *The Theory of Moral Sentiments,* III, 1, 2, p. 110.

[27] Smith, *The Theory of Moral Sentiments,* Volume 2.7, p. 204.

[28] Rawls, *Political Liberalism,* p. 12.

[29] Smith, *Lectures on Jurisprudence,* R.L. Meek, D.D. Raphael, and P.G. Stein, eds. (New York: Oxford, 1978), p. 104.

The relevance of distant perspectives has a clear bearing on some current debates in the United States, for example, that in the Supreme Court not long ago, on the appropriateness of using capital punishment for crimes committed in juvenile years. The demands of justice being seen to be done even in a country like the United States cannot entirely neglect the understanding that may be generated by how the problem is assessed in other countries in the world. The majority judgment of the Court did not simply "defer to like-minded foreigners" (as Justice Scalia suggested), but accepted that in the deliberations to arrive at grounded but nonparochial American judgments, it may be useful to take into account the enlightenment that nonlocal perspectives provide, after subjecting them to critical scrutiny in a better informed local framework.

Indeed, the apparent cogency of parochial values often turns on the lack of knowledge of what has proved feasible in the experiences of other people. The inertial defence of infanticide in ancient Greece could be influenced not only by the knowledge of societies in which infanticide is taken to be entirely unacceptable, but also from the fact that these societies are not crumbling into chaos and crisis as a result of not permitting such killing. Despite the undoubted importance of local knowledge, global knowledge has some value too, and can contribute to the debates on parochial values and practices.

DIVERGENCE AND THE FEASIBILITY OF THE COMPARATIVE APPROACH

In a model of arbitration there is likely to be some divergence of voices on a number of issues, especially when the perspectives sought come from far and wide. Even though there may be considerable convergence of values through global public reasoning, there might well be lasting differences between how the outcomes of such deliberations may emerge in different social and cultural settings. For a transcendental approach to justice dependent on the emergence of a complete agreement on the nature and demands of "global justice," this would of course be an overpowering problem. It is not, however, similarly problematic for the use of the comparative approach to justice. The focus here will be on whether there are significant issues on which agreements or consensus may emerge, especially after public interaction, with exchange of knowledge and understanding. The demands of global justice may not go beyond the agreed ways of enhancing justice in the world in the comparative route to judgments of justice.

Furthermore, the extent of agreement on global rankings may itself go on expanding as the process of interaction continues. The comparative approach does not require an "all or nothing" extremism, and it allows the world to come to grips with intense issues of global injustice (such as famines, widespread hunger, rampant illiteracy, or needless deaths from preventable or manageable diseases), on which consensus may be easier to obtain, without waiting for a full agreement on more contentious evaluations.

A similar thing can be said about taking note of global perspectives in making local decisions that are not primarily matters of global justice: for example, in assessing a national framework of punitive legislation, taking into account whether "a punishment appears equitable," to use Smith's phrase, without neglecting how it appears to people reared in a different background and with different informational understanding. The demands of objectivity not only require avoiding a "personal slant" (as Rawls noted), but also national parochialism (as Smith emphasized). This does not, however, yield any obligation to accept the views of others elsewhere, only that they be taken into account in an overall scrutiny (leading to modification of local priorities in some cases and no revision in others).

Indeed, distant voices too are subject to the discipline of critical scrutiny, including the invoking of impartial spectators from far as well as near. The discipline of fair arbitration in the context of global justice can be seen to be globally interactive. General *acceptability*, which must be distinguished from pre-existing ubiquitous acceptance, is an important issue in any social evaluation, and I have tried to discuss elsewhere why open and interactive public reasoning is centrally important for understanding the claims that human rights make, despite differences in manifest practices *between* countries, and also of course *within* each country.[30]

The common standpoint that may be seen to emerge on the basis of such associative scrutiny may be far from total, and the form of the concordance need not, in many cases, go beyond noting that some social arrangements are seriously unjust in a way that can be remedied, even though other comparisons may be hard to make without substantial ambiguity. A theory of justice need not turn up its nose against the vast reach and relevance of acceptable conclusions on human rights, or social justice, arrived at on the basis of public reasoning, even when they do not amount to a complete resolution of all the existing decisional issues about societal organization.[31]

---

[30] See my "Elements of a Theory of Human Rights," *Philosophy and Public Affairs*, XXXII (Fall 2004): 315–56. See also Joshua Cohen, "Procedure and Substance in Delib-

WHAT DIFFERENCE DOES IT MAKE?

I shall not try to summarize the paper but will briefly note some of the issues discussed here. First, I have argued for rejecting the tradition of focusing on the classic—and much invoked—question "What is a just society?" and in favor of concentrating instead on comparative questions of justice (linked to inquiries about advancing justice, or reducing injustice). I have presented some reasons for doubting that the transcendental question can typically be answered, but even if it could be, and a transcendental theory of justice were entirely successful in answering that classic question, it would not yield—directly or even indirectly—a comparative framework, which is needed for the actual assessment of justice. On the other side, a comparative theory of justice may be entirely viable and thoroughly usable without containing—or entailing—any answer to the grand question "What is a just society?"

Second, the specification of the demands of full justice, and of the elaborate institutional paraphernalia which have to be marshaled for the pursuit of Rawlsian modeling of justice, leaves open the question how we should assess whether some social change would advance the cause of justice or hinder it—questions that we constantly face in the world, within each country and in the field of global arrangements. For example, having decent patent laws about the production and distribution of pharmaceutical products may do a lot for the miserable and needlessly doomed AIDS patients in the world today, and arguments can be presented to suggest that this would reduce a

---

erative Democracy," in Seyla Benhabib, ed., *Democracy and Difference: Contesting the Boundaries of the Political* (Princeton: University Press, 1996), pp. 95–119; and Charles Beitz, "Human Rights as a Common Concern," *American Political Science Review*, xcv (June 2001): 229–82.

[31] I take up these issues more fully in a forthcoming book to be published by Harvard University Press, called *Freedom and the Theory of Justice*.

manifest injustice in the global society. But it will not, on its own, take us anywhere near the demands of Rawlsian transcendental justice.

Third, an implicit belief, which seems fairly common in a substantial part of political philosophy, that the identification of a fully just society is not only crucial for the comprehension of the nature of justice, but also essential for a well-founded relational understanding of justice, has had the effect of giving a fairly ubiquitous role to transcendental analysis of justice in contemporary philosophy. I have argued that this underlying belief may be entirely mistaken. A transcendental approach is neither necessary nor sufficient for answering questions on the advancement of justice that urgently demand our attention, which call for a robustly comparative approach. That approach is sharply different from the exploration of transcendence.

Fourth, there is a need to allow—and sometimes even to assert—incompleteness of relational comparisons of justice. Incompleteness can arise from unbridgeable gaps in information, but also from decisional unresolvability involving disparate considerations that may resist gradation, even with full information. However, possible incompleteness of judgments emerging from the relational route is not an embarrassment for practical reason. Indeed, far from it.[32] The challenge of assessing *advancement,* or identifying *regression,* will very often be not compromised at all by the presence of substantial incompleteness in the rankings of justice.

Fifth, aside from the general importance of incorporating possible incompleteness of evaluations, the admissibility of incompleteness also makes it easier to bring in distant voices in the assessment of justice, which can be critically important for the reach and strength of public reasoning. A procedural requirement to consider nonlocal perspectives can help to avoid undue dominance of local interest as well as possible parochialism of local reasoning shaped by the influence of established conventions and limited informational frameworks (without these being intellectually challenged). We have to go firmly beyond relying on the territorially moored perspectives of "free and equal citizens who are born into that society in which they lead their lives" (to use Rawls's phrase).[33]

The world in which we live is not only unjust, it is, arguably, extraordinarily unjust. It is not frivolous to seek a framework for a theory of justice that concentrates on advancement, not transcen-

---

[32] I have discussed this subject in "Consequential Evaluation and Practical Reason," and some of the underlying analytical concerns in "Maximization and the Act of Choice."

[33] Rawls, *Political Liberalism,* p. 23.

HAS GAME THEORY BEEN REFUTED?*

In his famous "Rational Fools" article, Amartya Sen[1] remarked that

> if today you were to poll economists of different schools, you would almost certainly find the coexistence of beliefs (i) that the rational behavior theory is unfalsifiable, (ii) that it is falsifiable and so far unfalsified, and (iii) that it is falsifiable and indeed patently false (*ibid.*, p. 325).

Three decades later, Sen's poll would give pretty much the same results. But although the confusion highlighted by Sen was not unjustified in the seventies, this is not the case anymore. Or so I will argue in this paper. This paper defends the view that game theory as commonly understood is not unfalsifiable, that it suffers from important recently discovered empirical anomalies, and that it cannot accommodate the anomalies in its theoretical framework.

A few clarifications are due before we start. First, I will be concerned with game theory in its *descriptive* (or "positive") interpretation only.[2] Second, asking whether game theory has been refuted does not imply a commitment to some naïve form of falsificationism. "Refuted theory" here is taken minimally to mean a theory that suffers from severe empirical anomalies. Finally, this paper completely sidesteps underdetermination and Duhem-Quine problems. It is not a paper about the methodological problems of empirical testing in

---

*The ideas in this paper have been presented at seminars at the University of Turin, the 2005 Economic Science Association meeting, and the 2005 British Society for the Philosophy of Science conference. A British Academy travel grant is gratefully acknowledged. I should thank the audiences at these meetings, Dario Castiglione, Jim Cox, John Dupré, Matthias Hild, Jim Woodward, Philippe Mongin, and an anonymous referee for their comments and suggestions. I was also inspired by several conversations with Dan Hausman and Chris Starmer. All the remaining mistakes are mine.

[1] "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs*, VI (1977): 317–44.

[2] On the interaction between normative and descriptive interpretations of game theory see Joseph Kadane and Patrick Larkey, "The Confusion of Is and Ought in Game Theoretic Contexts," *Management Science*, XXIX (1983): 1365–79. On the theory of individual decision making (in nonstrategic situations), see, for example, Francesco Guala, "The Logic of Normative Falsification: Rationality and Experiment in Decision Theory," *Journal of Economic Methodology*, VII (2000): 59–93; and Chris Starmer, "Normative Notions in Descriptive Dialogues," *Journal of Economic Methodology*, XII (2005): 277–89.

0022-362X/06/0305/239–63

general, but rather deals with the much more specific issue of game theory's relation with empirical evidence.

IS GAME THEORY A TAUTOLOGY?

Like other successful scientific theories, game theory can be (and is) interpreted in different ways. Thus first of all it will be necessary to clarify what we are dealing with: *What* has been refuted (if at all)? Under a standard interpretation, game theory is aimed at modeling and understanding people's behavior, in particular when the outcome of their actions depends not only on what they do, but also on other people's decisions. According to an influential textbook,[3] for example,

> Game theory is a bag of analytical tools designed to help us understand the phenomena that we observe when decision-makers interact. The basic assumptions that underlie the theory are that decision-makers pursue well-defined exogenous objectives (they are *rational*) and take into account their knowledge or expectations of *other* decision-makers' behavior (they reason *strategically*) (*ibid.*, p. 1).

This characterization rules out, first of all, so-called "mass-action" interpretations of game theory, as for instance in evolutionary game theory models. Game theory, under the standard "epistemic" interpretation, models individual behavior as determined by players' preferences and beliefs, under strong rationality assumptions. This does not mean that mass-action interpretations are less interesting or promising than the epistemic one, but simply that the latter is by far the most developed, used, and—crucially, for our purposes—empirically tested version of game theory.[4]

The standard textbook definition also rules out interpreting game theory as a branch of pure mathematics. Although it is common to distinguish "pure" from "applied" game theory, the former is *not* a piece of uninterpreted formalism. The Osborne and Rubinstein quote significantly mentions decision makers, expectations, knowledge, objectives, and behavior, as the subject matter of game theory. Of course there is no mention of what these objectives may be, nor indeed of what the expectations are about. "Decision makers" is also left vague enough to be applicable to various real-world entities (individuals, firms, nations). But this just means that "pure" game theory is *very abstract*, rather than purely formal.[5] The business of specifying preferences, beliefs, and the identity of decision makers is left to applied social scientists, and presumably varies depending on the intended domain application.

---

[3] Martin Osborne and Ariel Rubinstein, *A Course in Game Theory* (Cambridge: MIT, 1994).

[4] The distinction between these two interpretations goes back to John Nash's seminal Ph.D. dissertation.

There are some dissenters, however. "Pure" game theorists often claim that their theory is just a *tautology*. Ken Binmore[6] is a typical example:

> Mathematical theorems are tautologies. They cannot be false because they do not say anything substantive. They merely spell out the implications of how things have been defined. The basic propositions of game theory have precisely the same character (*ibid.*, p. 96).

As with other claims of this kind, it is useful to ask first what is *meant* by "tautology," and then *what* is supposed to be a tautology.[7] A tautology is, strictly speaking, a proposition that is true in virtue of the definition of its logical terms. This is probably not what Binmore has in mind. Most likely, he is thinking of *sets* of propositions such as:

(1) Rational agents always choose uniquely dominant strategies.
(2) "Defect" is the only dominant strategy in the one-shot Prisoner's Dilemma game.
(3) Therefore, rational agents always defect in the one-shot Prisoner's Dilemma game.

This is an argument, not a proposition. It is a valid argument, in the standard logical sense: if the premises are true, the conclusion must also be true, independently of the truth of the premises. But whether each proposition in the argument is true or false is another question altogether. In order to figure out whether it is true that rational agents always defect in the one-shot Prisoner's Dilemma game, we need to check whether rational agents always choose uniquely dominant strategies, and whether "defect" is the only dominant strategy in the one-shot Prisoner's Dilemma.[8]

---

[5] A similar point could be made (more elaborately, perhaps) by looking at the sort of solution concepts that are explored and considered legitimate in game theory. Rationality plays a prominent role in this activity, and rationality is not a purely formal concept. It is rationality for decision makers, and more specifically decision makers of the *human* kind.

[6] *Game Theory and the Social Contract, Volume 1: Playing Fair* (Cambridge: MIT, 1994).

[7] See, for example, Elliott Sober, *Philosophy of Biology* (Boulder: Westfield, 1993), pp. 68–72.

[8] On this point, see also Daniel Hausman, "'Testing' Game Theory," *Journal of Economic Methodology*, XII (2005): 211–23.

Table 1.

	Cooperate	Defect
Cooperate	2, 2	0, 3
Defect	3, 0	1, 1

But how do we check that (1) and (2) are true? Binmore and the other supporters of the "tautology" interpretation presumably argue that (1) and (2) are *analytic*: they are true in virtue of the meaning of the nonlogical terms that are in them. So "rational agent" is defined as the kind of decision maker that always chooses uniquely dominant strategies. And the one-shot Prisoner's Dilemma game (Table 1)[9] is by definition a game where "defect" is the only dominant strategy.

Of course there would be nothing to test, if game theory was nothing but this set of propositions. "Tautologists" like Binmore however do not conclude that game theory is empirically useless:

> One reason for emphasizing the tautological nature of these tools of the game theory trade is to direct attention of the critics away from the question of how a formal game is *analyzed* to the more rewarding question of how problems of human interaction should be *modeled* as formal games (*ibid.*, p. 169).

So what can be tested presumably is the claim that certain game theory models can be used to understand (predict, explain) certain specific real-world situations.[10] This seems analogous to what some versions of the semantic view of scientific theories call a "theoretical hypothesis" (but perhaps the expression "empirical hypothesis" would be less misleading).[11] A testable hypothesis would be, for example, that the one-shot Prisoner's Dilemma game can be used to explain (predict, and so on) the behavior of experimental subjects playing the game in Table 2.

---

[9] In representing games in normal (or strategic) form, I follow the usual conventions: the first number in each cell (or consequence, or outcome) represents the payoff of the "Row" player, the second number the payoff of the "Column" player. Each row/column represents a strategy (or act) available to that particular player.

[10] See also Binmore, "Why Experiment in Economics?" *Economic Journal*, CIX (1999): F16–24, p. F18: "[a consolidating experimentalist] does not see as his task as asking whether economics works or not. He already knows that sometimes it does and sometimes it does not. The task is to classify economic environments into those where the theory works and those where it does not."

[11] See, for example, Ronald Giere, *Explaining Science* (Chicago: University Press, 1988), chapter 3.

Table 2.

	Left	Right
Up	$2, $2	$0, $3
Down	$3, $0	$1, $1

So far it looks as if the tautologists are just quibbling about what a scientific theory is. According to them, a theory is just the formalism. According to most scientists and philosophers, it is an *interpreted* formalism, or a formalism plus a set of "bridge principles" that connect it to empirical reality. But in fact some tautologists are more radical than that. Take Binmore again: "Game theorists understand the notion of a payoff in a sophisticated way that makes it tautologous that players act as though maximizing their own payoffs."[12] This strategy is defended by drawing an analogy with biology:

> [in applications of game theory in evolutionary biology] the payoffs are...defined in terms of an organism's "fitness". This makes it a tautology that the organisms that survive will seem to behave as though seeking to maximize their payoffs in the game under study. The same tautological flavor is also to be found when *rational* behavior in games is studied, provided that the payoffs in the game are interpreted according to the theory of *revealed preference...* (*ibid.*, p. 104).

Revealed preference theory was originally proposed as an attempt to *reduce* the notion of preference to that of observed choice.[13] There is remarkable disagreement nowadays about what revealed preference theory really amounts to in neoclassical economics,[14] but tautologists like Binmore seem to stick close to the original understanding: "In revealed preference theory...choice behavior becomes the primitive. One observes some of the choices that a player makes and then argues that he is making choices *as though* he were equipped with a preference relation" (*ibid.*, pp. 105–06).

The first thing to notice is that Binmore's analogy with fitness in evolutionary biology is far from unproblematic. According to a common interpretation, the fitness of a trait X is not whatever proportion of offsprings with X are observed in the next generation. Fitness is the *expected* reproductive success of a trait, or the probability—given the forces of natural selection—that a member of the next generation of offsprings will have trait X. Probability here can be (and usually is) interpreted as a *propensity*, rather than a frequency. This probability may or may not be reflected approximately in the frequency of X in the next generation. It may not be because other factors besides natural selection can interfere with the heritability of X: random genetic drift, typically, or mutation, migration, recombination. It is not true that "We know who will survive because they are fittest" and "We know who are fittest because they survive" (*ibid.*, p. 98).

---

[12] *Game Theory and the Social Contract*, p. 98.

[13] Paul Samuelson, "A Note on the Pure Theory of Consumer's Behavior," *Economica*, v (1938): 61–71.

[14] See, for example, Daniel Hausman, "Revealed Preference, Belief, and Game Theory," *Economics and Philosophy*, xvi (2000): 99–115; Keith Dowding, "Revealed Preference and External Reference," *Rationality and Society*, xiv (2002): 259–84.

Interpreted this way, fitness is not a "tautology." This of course does not, by itself, prevent one from interpreting utility in terms of revealed preferences. But there are other, independent reasons not to do so. Preferences cannot be "read off" directly from behavior because choices are determined by individual preferences *and beliefs*. This point is perhaps obscured by the convention in game theory of assuming common knowledge of the game that is being played ("I know that you know that I know...(and so forth) that such and such strategies are available, with such and such payoffs, and that all players are rational"). Of course *if* we assume common knowledge of the game, then we can derive preferences from observed choices. Whether the derivation is correct, however, depends crucially on the correctness of the common knowledge assumption.

It is not clear however why one should want to take such a strong empiricist stance regarding preferences, if one is willing to simply assume a priori a certain configuration of players' beliefs.[15] Why accord to the two key variables of game theory a different status? In fact the fathers of rational choice theory were fully aware of the need *mutually* to determine, by means of observable procedures, the structure of individual preferences *and* beliefs.

SAVAGE MEASUREMENTS

I will keep using an old-fashioned philosophical distinction, according to which a scientific theory can be analyzed in two sets of elements. On the one hand, a set of abstract concepts connected by theoretical relationships (or "internal principles"). On the other, some "bridge principles" that link the concepts with empirical reality or measurement procedures.[16] What are the bridge principles of game theory? In *The Theory of Games*, John von Neumann and Oskar Morgenstern[17] famously introduce a procedure for measuring (or "operationalizing") cardinal utility. Expected utility theory is the "bridge" connecting game theory with the real world.

---

[15] See, for example, Hausman, "Revealed Preference," pp. 101–06.

The story has become part of the folklore of game theory. Von Neumann was happy to build the theory on the generic notion of "payoff." Indeed, he conceived of payoffs basically as *objective* (material, possibly monetary) properties of outcomes or consequences. Morgenstern persuaded von Neumann to recast the theory in utility jargon, in order to appeal to the average neoclassical economist. Von Neumann quickly scribbled on the back of an envelope the measurement procedure that was to appear in the second (1947) edition of *The Theory of Games.*

Von Neumann and Morgenstern's measurement procedure of utility is well known: starting with three outcomes x, y, and z such that $x > y > z$, the utility scale is normalized by assigning $U(x) = 1$ and $U(z) = 0$. A subject is then presented with the choice between the prospect y for sure and the prospect $[p, x; (1 - p), z]$; if she prefers either the risky prospect or the sure outcome, the probabilities in the former are varied until she is indifferent. At this stage, the utility of y can be computed by $U(y) = p$, and the procedure iterated for any value between x and z.

This clearly assumes that we already know subjects' probability assignments or beliefs. In practice, since von Neumann and Morgenstern restrict their discussion to lotteries with objective probabilities, we just assume that subjects' assignments mirror the objective probability measures (their beliefs are accurate representations of the world). When this is not the case, subjective probability assignments must be worked out from people's choices, in a way that is symmetric to the one above. Start with people's preferences about sets of consequences, and then infer which states of the world are considered more probable from the observation of individuals' acts (or choices of strategy). In practice, it is usually assumed that more money is preferred to less, and then individuals are offered money prizes associated with certain types of event. But again one has to take a firm fulcrum (preferences for the measurement of beliefs, beliefs for the measurement of preferences) for the lever of expected utility theory to lift any weight.

---

[16] We do not have to assume that these two elements are precisely separable. The distinction is useful only for heuristic purposes, and should not be read as implying any deep philosophical commitment on the theory-observation distinction, or on the nature of scientific theories. In fact, it should be broadly compatible both with old-fashioned "syntactic" approaches (like the "standard view" of theories) and with more fashionable "semantic" approaches (where the "bridge principles" are sometimes called "theoretical hypotheses," see above).

[17] *The Theory of Games and Economic Behavior* (Princeton: University Press, 1944).

These procedures were first sketched by Frank Ramsey and then refined by Leonard Savage.[18] Since the latter's systematization has become standard in contemporary game theory, I shall refer to it as the *Savage measurement procedure.* According to "kosher" game theory, the application of the theory should start with a Savage measurement, that is, with the identification of players' utilities and beliefs. Once this has been done, one can begin to figure out what kind of game they are playing, and apply the solution concepts of game theory.[19] Most experimental game theory is not kosher from this respect, however. This causes problems, because the results of experiments can be (and are) criticized for not testing the theory at all.

WHAT KIND OF GAME ARE WE PLAYING?

Take the one-shot Prisoner's Dilemma game again (Table 1 above). Here the free riding strategy is Down for Row and Right for Column, leading to a payoff of 1 unit each. These strategies are uniquely dominant, and determine the unique rational solution (the Nash solution or Nash equilibrium) of the game. Informally, whatever Column will do, Row will be better off by playing Down; and similarly Column will always be better off by playing Right, given what the opponent does. But both would be better off if they played Up-Left. The Nash equilibrium in a one-shot Prisoner's Dilemma game is not a Pareto-optimal solution.

---

[18] Ramsey, "Truth and Probability," in Richard Braithwaite, ed., *The Foundations of Mathematics and Other Logical Essays* (New York: Routledge, 1931), pp. 156–98; Savage, *The Foundations of Statistics* (New York: Wiley, 1954).

[19] Sometimes Binmore writes (and speaks, in conversation) as if his revealed preference theory was nothing but an application of Savage's measurement procedure. He says for example ("hawk" here stands for "defect," "dove" for "cooperate"): "the preference (*dove, hawk*) $\prec_A$ (*hawk, hawk*) is deduced from the prior information that, if Adam knew he had to choose between only (*dove, hawk*) and (*hawk, hawk*), then he actually would choose (*hawk, hawk*)" (*Game Theory and the Social Contract*, p. 106). Notice that according to this formulation one does *not* observe behavior in a Prisoner's Dilemma game first, and *on the basis of this,* define the preference scales. Rather, one elicits the preference scales by pair wise comparison of the various outcomes (as in a simplified Savage measurement—more about this below), and *then* uses such information to predict behavior in the Prisoner's Dilemma game. This does not make a "tautology" out of game theory, because of course there may well be a mismatch between the preferences as "revealed" in the two situations. And it does not reduce the concept of preference to that of choice, because in order to identify preferences in the simplified Savage procedure you need to postulate a given belief structure.

Table 3.

	Left	Right
Up	2, 2	0, 0
Down	0, 0	1, 1

As is well known, in a "standard" Prisoner's Dilemma experiment a considerable number of subjects play cooperatively (Up-Left).[20] The game played in such experiments, however, is not necessarily the one in Table 1. It is the one already presented in Table 2 above. The "naive" interpretation of these experimental results is that many human beings (fortunately) do not behave as predicted by the theory. But a number of game theorists reject this interpretation. They say, for example, that the preference rankings of the subjects who play cooperatively in these experiments are inadequately represented by the numbers in the classic Prisoner's Dilemma matrix. Or, in other words, that Table 1 and Table 2 are not necessarily isomorphic for all experimental subjects.

According to the orthodox interpretation of game theory, the numbers in Table 1 represent the (ordinal) structure of agents' preferences. Thus, the argument goes, if we observe anomalous behavior in the experiment, then the initial conditions postulated in the model were probably not instantiated in the experiment. Subjects were not *really* playing the Prisoner's Dilemma, but another game of their choice. A game like the one in Table 3, for example.

This game has *two* Nash equilibria. The transformation of Row's payoff from 3 to 0 in the Down-Left cell can be explained, for example, by Row's anticipated feelings of guilt in letting Column down, given that the latter was willing to cooperate (and vice-versa for Column in Up-Right). The Prisoner's Dilemma is then transformed in a coordination game, where Up-Left is a legitimate equilibrium.

Binmore has been a vocal supporter of this interpretation:[21]

> [those who are anxious to deny that people seek only their own narrowly conceived selfish ends] argue that the players may care about the welfare of their opponents, or that they may actively want to keep their promises out of feelings of group solidarity or because they would otherwise suffer the pangs of a bad conscience. Such players will *not* be playing the

---

[20] See, for example, Anatol Rapoport and Albert Chammah, *Prisoner's Dilemma: A Study in Conflict and Cooperation* (Ann Arbor: Michigan UP, 1965).

[21] *Fun and Games: A Text on Game Theory* (Lexington, MA.: D.C. Heat, 1992).

> Prisoners' Dilemma. They will be playing some other game with different payoffs (*ibid.*, pp. 313–14).

In other words: it is a common fallacy to interpret the payoffs in a game theory model as representing physical properties or (typically) monetary outcomes. They represent utilities. And people may derive utility from (or, more precisely, may have preferences about) whatever features of the consequences they happen to care about—including other players' payoffs, one's own anticipated guilt, regret, and so forth.

This is in line with the liberal tradition that informs much of economic theory. According to this tradition, it is important that social scientists remain neutral about the contents of people's tastes and beliefs. From a normative viewpoint (and economics is a theory of rationality as well as a theory of actual behavior) the requirements imposed on preferences and beliefs should be purely formal, of the *consistency* type. (Remember Hume: it is perfectly rational to prefer the end of the world to a scratch on my finger.) So, even though economists often assume for simplicity that outcomes are defined strictly in monetary terms, this ought not to be always the case: outcomes can (and should, whenever required) be redescribed or "refined" to include whatever properties the agents find relevant.

This idea (the possibility of "refining" the consequences) has provoked controversy since the early days of expected utility theory.[22] An explicit attempt to legitimize it is due to Peter Hammond, who puts the "refining" strategy right at the core of his "consequentialism." Consequentialism, as usually understood, is the idea that rationality is forward-looking and independent of procedural considerations.

---

[22] A (probably incomplete) list of contributions to this debate includes Paul Samuelson, "Utility, Preference and Probability," in Joseph Stigliz, ed., *The Collected Papers of Paul Samuelson, Volume 1* (Cambridge: MIT, 1952), pp. 127–36; Amos Tversky, "A Critique of Expected Utility Theory: Descriptive and Normative Issues," *Erkenntnis*, IX (1975): 163–73; Sen, "Rationality and Uncertainty," *Theory and Decision*, XXII (1985): 109–27, and "Internal Consistency of Choice," *Econometrica*, LXI (1993): 495–521; Mark Machina, "Dynamic Consistency and Non-Expected Utility Models of Choice Under Risk," *Journal of Economic Literature*, XXVII (1989): 1622–68; Michael Bacharach, "Preferenze Razionali e descrizioni," in Maria Carla Galavotti and Guido Gambetta, eds., *Epistemologia ed economia* (Bologna: CLUEB, 1988), pp. 43–63, Peter Hammond, "Consequentialist Foundations for Expected Utility Theory," *Theory and Decision*, XXV (1988): 25–78; Robert Sugden, "Rational Choice: A Survey of Contributions from Economics and Philosophy," *Economic Journal*, CI (1991): 751–85; John Broome, *Weighing Goods* (Cambridge: Blackwell, 1991); Paul Anand, *Foundations of Rational Choice Under Risk* (New York: Oxford, 1992); and Bertrand Munier, "Comment," in Kenneth Arrow, Ennio Colombatto, Mark Perlman, and Christian Schmidt, eds., *The Rational Foundations of Economic Behavior* (New York: St. Martin's, 1996), pp. 43–47.

> Consequentialism requires everything which should be allowed to affect decisions to count as a relevant consequence—behavior is evaluated by its consequences, and nothing else (*op. cit.*, p. 26).

Hammond's "consequentialism" tries to transform genuine consequentialism into an entirely general but admittedly trivial proposition:

> If regrets, sunk costs, even the structure of the decision tree itself, are relevant to...behavior, they are therefore already in the consequence domain. Indeed the content of a...theory of behavior is then largely a matter of what counts in practice as a relevant consequence, rather than whether consequentialism and other abstract axioms are satisfied (*op. cit.*, p. 26).

Many people find this problematic. It is often argued that if you are allowed to redefine consequences (or the argument of the utility functions) as you like, then economic theory becomes empty. This critique goes back a long time,[23] but is essentially misguided. It is a false dilemma to stipulate that *either* the payoffs are to be interpreted as money, *or* the theory is empty. People have preferences, and they are preferences *about* something. There is no reason a priori to assume that all individuals care about the same things. One difficult empirical task in testing game theory is precisely to figure out what they care about. Once we know what they care about, we can check whether they are rational—that is, whether the models of rational choice provide a correct account of human behavior. The theory is not empty. It can be used to say something substantial about the real world.

To illustrate, consider an analogy with classical mechanics. Newton's law of universal gravitation ($F = ma$) can be interpreted (and usually is interpreted) as a description of the relation between mass, acceleration, and force once *all* the forces at work in a given system have been taken into account. As such, it seems hard to falsify. Imagine a fictional classical (pre-relativity) physicist observing a prima-facie violation of the law; before coming to any conclusion about the validity of the universal law of gravitation, she would begin to search for some unobserved forces at work on the system under study. Such forces could be gravitational, but also of some other kind, not (yet) described by Newtonian mechanics (electromagnetic forces, for example). But the physicist's behavior would not make the law empty:

---

[23] See, for example, Terence Hutchison, *The Significance and Basic Postulates of Economic Theory* (New York: Kelley, 1938), pp. 114–15; for recent versions, see Machina, "Dynamic Consistency," pp. 83–87; Alexander Rosenberg, *Economics: Mathematical Politics or Science of Diminishing Returns?* (Chicago: University Press, 1992), chapter 5; and Anand, *Foundations of Rational Choice Under Risk*, pp. 105–06.

the law says that *once* all the forces have been taken into account, they obey the universal law of gravitation—and in some cases we might well be wrong in computing the total force in the system.[24]

So, there is a division of labor in classical mechanics: on the one hand general principles like the law of universal gravitation are supposed to hold for all bodies and all forces. On the other, more specific laws connect specific kinds of forces to other quantities of interest. For example: gravitational forces are described by equations like $F=G(m_1m_2/r^2)$; electrostatic forces by Coulomb's law ($F=q_1q_2/4\pi\varepsilon_0r^2$), and so on. Similarly in economics one can envisage a division of labor between general principles of rational play like those of game theory, and the more specific task of connecting payoffs (utilities) with empirical reality, by specifying the argument of the utility function.

Game theorists like James Cox and Jorgen Weibull have expressed this viewpoint clearly in recent articles.[25] Here is Cox, for example:

> In their seminal work on game theory, von Neumann and Morgenstern (1944, 1947) thought it necessary to simultaneously develop a theory of utility and a theory of play for strategic games. In contrast, much subsequent development of game theory has focused on analyzing the play of games to the exclusion of utility theory. In the absence of a focus by game theorists on utility theory, it is understandable that experimentalists testing the theory's predictions have typically assumed that agents' utilities are affine transformations of (only) their own monetary payoffs in the games. This interpretation of game theory incorporates the assumptions that agents do not care about others' (relative or absolute) material payoffs or about their intentions. There is a large experimental literature based on this special-case interpretation of the theory, which I shall subsequently refer to as the model of "self-regarding preferences." The part of the literature concerned with public goods experiments and trust and reciprocity experiments has produced replicable patterns of inconsistency with predictions of the model of self-regarding preferences. For example, the patterns of behavior that have been observed in one-shot trust and reciprocity games are inconsistent with the subgame perfect equilibria of that model. But this does *not* imply that the observed behavior is inconsistent with game theory, which is a point that has not generally been recognized in the literature (*op. cit.*, pp. 260–61).

---

[24] This fictional example is not far from historical reality, as shown by Imre Lakatos, "Falsificationism and the Methodology of Scientific Research Programmes," in *Philosophical Papers, Volume 1* (New York: Cambridge, 1970), pp. 8–101, and others.

[25] Cox, "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, XLVI (2004): 260–81; see also Weibull, "Testing Game Theory," in Stephan Huck, ed., *Advances in Understanding Strategic Behavior* (New York: Palgrave, 2004), pp. 85–104, especially pp. 85–86.

The message is: one cannot proceed to make any inference about the empirical status of game theory without figuring out first what individuals care about. If they care about fairness, for example, it is still entirely possible that they are rational maximizers of utility—once the latter has been defined appropriately to capture fairness considerations. The definition of the utility function as ranging exclusively over one's own monetary outcomes is unnecessarily restrictive and in principle can be relaxed to allow for more varied (and interesting) individual preferences.

PROBLEMS WITH REFINEMENTS

The possibility of refining the options does not make the theory empty. There are still two ways in which game theory can be refuted. First, suppose that cooperators in the one-shot Prisoner's Dilemma really did care about their own monetary outcomes only (and preferred more money to less). Perhaps they do not understand the concept of dominance; perhaps they do not pay enough attention to the structure of the game. If either of these hypotheses were correct, the theory would turn out to be false.

This is just a speculation. As a matter of fact most experimental game theorists (including Cox and Weibull) agree that to assume selfish preferences over money only is too restrictive. There is more in the argument of our utility functions, and this extra content should be accommodated in the framework of game theory.

This leads to the second way in which game theory could turn out to be refuted. Suppose that subjects cared about something else besides their own monetary gain, but that game theory was not flexible enough to accommodate this "something else" within its framework. Were the bridge principles unable to account for the observations, in particular, the only option left would be to modify the theory of rational play (the "hard core" of game theory).

This possibility is rarely discussed, perhaps because of a taken-for-granted analogy between folk psychology and formal rational choice theory. According to our everyday understanding, *any* observed pattern of behavior can in principle be rationalized by imputing an appropriate—perhaps highly sophisticated—structure of beliefs and desires. The intertemporal coherence of beliefs and desires can also always be restored by simply modifying their contents—no matter what the individual does or is going to do next. Alexander Rosenberg is the philosopher who has most forcefully argued that "the fundamental explanatory strategy of economic theory is of a piece with that of our ordinary explanations of human action" (*op. cit.*, p. 118). And the view that "no intentional action can be internally irrational" (in

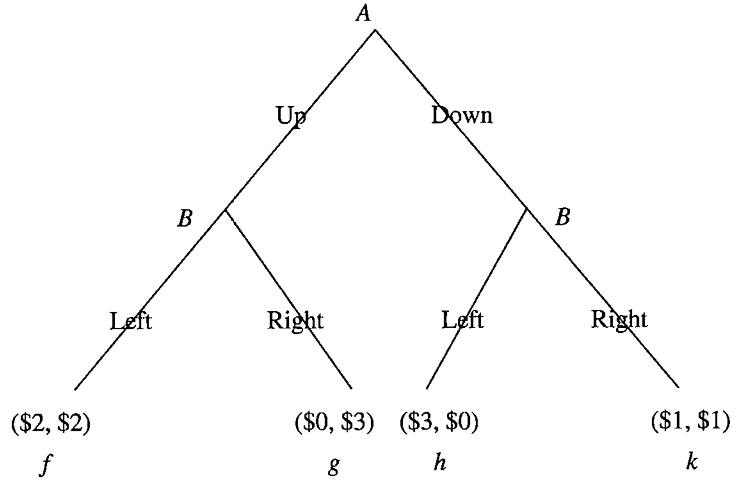


Figure 1

the above sense), has been dubbed the *Plato principle* by Donald Davidson.[26]

The Plato principle, however, is false when applied to game theory, because the formal theory of rational choice is less flexible than folk psychology, and as a consequence has more limited explanatory resources. This applies to the theory both in its normative and in its descriptive version, but given the focus of this paper, let us phrase it in a descriptive idiom. There is a tension between certain refinements of the consequences and the Savage measurement procedure. Consider the sequential Prisoner's Dilemma game in Figure 1.[27] The first monetary prize between brackets at each end-node is *A*'s monetary payoff, the second one is *B*'s monetary payoff; the letters *f*, *g*, and so on, are arbitrary labels assigned to the consequences. Experimenters report data where 40% of first-movers (*A*) play a *trust* move (Up) in a game of this kind, and second-movers (*B*) respond Left 35% of the time.[28]

Can we "refine" *A* and *B*'s preferences in order to account for this behavior? It depends. Suppose *A* and *B* are motivated by two sets of considerations. They prefer more money to less, but also have a taste for fairness. In equation (1) deviations from fairness are defined with respect to a standard "fair payoff" π^f:

---

[26] See, for example, his "Paradoxes of Irrationality," in Richard Wollheim and James Hopkins, eds., *Philosophical Essays on Freud* (New York: Cambridge, 1982), pp. 289–305, see p. 294.

[27] Borrowed from Hausman, "Revealed Preference," p. 107.

[28] See, for example, Kenneth Clark and Martin Sefton, "The Sequential Prisoner's Dilemma: Evidence on Reciprocation," *Economic Journal*, CXI (2001): 51–68.

$$(1)\ V_i = \pi_i - \alpha_i(|\pi_i^f - \pi_i| - \sum_{j\neq i}^{n} \beta_i|\pi_j^f - \pi_j|).$$

Player i's material payoffs are denoted by π_i, and the parameter α introduces i's concern for the pattern of distribution of the payoffs (as opposed to his own payoffs). When α=0, the player is purely selfish (his utility V_i is a direct function of his material payoff only). Everything on the right-hand side of α represents fairness considerations. When the standard of fairness is equality of payoffs ($\pi_i^f = \pi_j^f$), this becomes a model of inequality aversion.[29] The parameter β weighs the impact of the first set of fairness considerations (inequality in one's payoffs) against that of the second one (inequality in other players' payoffs).

This is just one example, representative of a wider class of models of fairness preferences. It suffices to raise two points. First, these are not alternatives to the theory of rational play at the core of game theory. They are attempts to pin down the bridge principles connecting game theory with observed behavior (or better: with observable features of the outcomes, like the distribution of monetary gains π_i). Such models play in the hands of the refiners, and their empirical successes in predicting experimental data are not failures of game theory.

Second, these models add to the refiners' weaponry because they are *genuinely* consequentialist in spirit. The extensive form of the game in Figure 1 is quite redundant. The equation in (1) can be applied simply by looking at the normal form of the game (Table 2 above). In principle, this is fully compatible with the orthodox Savage measurement procedure. Indeed, equation (1) could be read as a generalization of some class of preference structures observed by means of the Savage procedure.

How would this measurement work, in practice? Suppose we want to measure *B*'s utilities. All we have to do is to present player *B* with the choices in Figure 2. *B*'s preferences, as revealed in these binary-choice tasks, will suffice to predict her behavior in the sequential

---

[29] See for example, David Levine, "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, I (1998): 593–622; Ernst Fehr and Klaus Schmidt, "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics*, CXIV (1999): 817–68; Gary Bolton and Axel Ockenfels, "ERC: A Theory of Equity, Reciprocity and Cooperation," *American Economic Review*, XC (2000): 166–93.

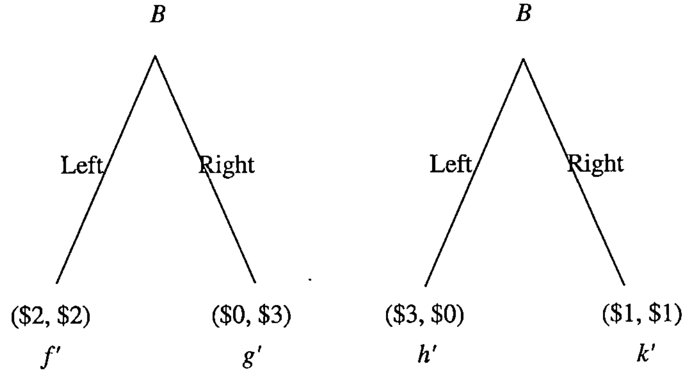


Figure 2

Prisoner's Dilemma game, *if* they conform to the preference structure of (1). In this scenario, $f=f'$, $g=g'$, $h=h'$, and $k=k'$.

But let us suppose instead that *B* is not primarily motivated by distributional concerns. The reason she plays the equal payoff (*f*) in Figure 1 could be that she is *reciprocating* *A*'s first move. She chooses Left in response to a "trust" move because she wants to be kind towards *A*, but has no reason to do so when *A* has not played at all. In this scenario, looking at *B*'s choices in Figure 2 will be useless.[30]

MODELING RECIPROCITY

Reciprocity is a tricky notion. Consider the *basic reciprocity formula*:

$$(2)\ V_i = \pi_i + \rho[\kappa(.)\sigma(.)]$$

Here ρ is a parameter weighing the impact of reciprocity considerations against the usual self-interested preference for one's own monetary gains (π_i). κ is a *kindness function*, representing how "nice" or "offensive" *j*'s behavior is towards *i*. The other function, σ, captures *i*'s reaction to *j*'s behavior. A *negative* kindness combined with a *positive* reaction (or vice-versa) results in a reduction of overall utility (V_i). Reacting negatively to negative kindness, and positively to positive kindness, leads to an increase of overall utility.

At first sight, it may look as if the arguments of these two functions could be filled in consistently with the spirit of genuine consequentialism. But this is unlikely. Consider the reaction function first: a negative reaction is one that *hurts j*, typically by reducing her payoff with respect to a given standard or reference point. The standard is most naturally interpreted as what *j would have* achieved, had *i* played a different strategy. We shall represent this possible payoff with π_j'. Similarly, *j*'s kindness towards *i* depends not only on *i*'s relative payoff with respect to some absolute standard (say, equality, as in most models of fair preferences), but on what *i could* have achieved, had *j* played differently. Again, we will use π_i' to represent such counterfactual payoff, where the counterfactual mark (′) refers to the fact that it is a *possible* payoff—an outcome that could in principle be achieved in *this* game.[31]

---

[30] See also Hausman, "Revealed Preference," p. 108, for this point.

$$(3)\ V_i = \pi_i + \rho[\kappa(\pi_i - \pi_i')\sigma(\pi_j - \pi_j')]$$

The most advanced models of reciprocity in this spirit make use of the tools provided by *psychological game theory*.[32] In psychological game theory the payoffs depend not just on the material structure of the consequences, but also on players' *beliefs*. Following Matthew Rabin's seminal work,[33] reciprocity can be represented as involving a calculation of what player *i* believes that *j* believes that *i* will play. Player *j*'s intention is perceived as "nasty" when *i* believes that *j* believes that *i* will make a "nice" move, and yet *j* chooses a strategy that reduces *i*'s monetary payoffs with respect to the reference point.[34]

A discussion of psychological game theory is besides the scope of this paper. How counterfactual considerations are modeled is not crucial: perhaps intentions are not even relevant *all* the time; perhaps in some cases *B* reacts to *A*'s "nasty" moves only to signal that she is making a mistake, or in order to push the outcome towards a Pareto-superior solution (*B* might be driven by selfish motives, that is, and might have a crude stimulus-response model of *A*'s behavior).[35]

---

[31] I am simplifying a lot here. The counterfactual "reference point" in $k(\pi_i - \pi_i')$ is in many occasions better represented as $\pi_i'^f$, where *f* stands for the "fairest" payoff that could have been achieved in the game. It is likely that $\pi_i'^f$ should be identified in each game by comparing *i* and *j*'s various possible payoffs.

[32] First proposed by John Geannakoplos, David Pearce, and Ennio Stacchetti, "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, I (1989): 60–79.

[33] "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, LXXXIII (1993): 1281–302.

[34] The two prominent models in this tradition are Armin Falk and Urs Fischbacher, "A Theory of Reciprocity," Working Paper 6 (2000), Institute for Empirical research in Economics, University of Zurich; and Martin Dufwenberg and Georg Kirchsteiger, "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, XLVII (2003): 268–98.

[35] As a matter of fact, there is some puzzling evidence suggesting that in dilemma experiments some subjects "reciprocate" even when they are playing against a computer; see Daniel Houser and Robert Kurzban, "Revisiting Kindness and Confusion in Public Goods Experiments," *American Economic Review*, XCII (2002): 1062–69.

What really matters is the role of counterfactual considerations themselves. The identification of the preferences associated with each consequence requires an examination not just of what will happen, but also of *what might have happened.* To put it another way, the utilities are context-dependent: they do not just depend on the outcomes taken in isolation, but on the *whole structure of the game.*

RECIPROCITY CANNOT BE "REFINED"

Remember Hammond: "Consequentialism requires everything which should be allowed to affect decisions to count as a relevant consequence"; if reciprocity affects decisions, it must be accounted for at the moment of describing the consequences, according to consequentialism.

This turns out to be problematic in the Savage framework, however. The problem has been around for a long time. In a letter from Robert Aumann to Savage, in 1971,[36] the issue is first raised of how to distinguish *acts* (say, taking an umbrella) from *consequences* (for example, getting wet). Aumann points out that the distinction is fuzzy. Savage agrees but says in a very pragmatic vein that this may not matter as long as "I seem to be able to couch my decision problems in terms of them"; for "...a consequence is in the last analysis an experience" (*ibid.*, p. 79).

But this is overoptimistic. Savage defines *acts* as functions from *states of the world* to *consequences.* (For example: the act "taking the umbrella" assigns the consequence "staying dry" both to the state "it rains" and to the state "it does not rain"; "not taking the umbrella" assigns the consequence "getting wet" to the state "it rains" and the consequence "staying dry" to the state "it does not rain.") All the existing versions of expected utility theory prove the existence of the utility function (representation theorem) from the assumption that decision makers form preferences over *every* act that can possibly be constructed by combining consequences with states of the world. John Broome[37] calls this feature the *rectangular field assumption.*

The rectangular field assumption is not *implied* by expected utility theory, from a logical point of view. But the existence of the expected utility function is derived from the rectangular field assumption and other fundamental principles.[38] So a violation of the rectangular field assumption does not logically contradict expected utility theory (although we still lack an alternative representation theorem that does without that assumption). It does, however, seriously invalidate the Savage measurement procedure.

---

[36] Reprinted in Jacques Drèze, *Essays on Economic Decisions Under Uncertainty* (New York: Cambridge, 1987), pp. 76–81.

[37] Broome, *Weighing Goods*, pp. 115–17; see also Sugden, "Rational Choice," pp. 761–63.

[38] Broome, *Weighing Goods*, p. 117.

For measurement purposes it is crucial that acts can be constructed arbitrarily. Take any state of the world, and combine it with any consequence in our theoretical universe: whatever we obtain by these arbitrary combinations must make sense, and must be something the agent can form a preference about. Now suppose that the consequence *g* of player *B* in the sequential Prisoner's Dilemma game of Figure 1 is defined as "getting $3 but being sorry for player *A*" (who had offered to cooperate). This is surely meaningful in the context of a game like the one above. But does it make sense in a different context, where that consequence does *not* follow from the act of choosing Right in the above game (after *A* has chosen Up)?

We must distinguish two cases (*ibid.*, p. 116):

(1) The (refined) consequence is *causally implausible* outside the specific game; for example, why should *B* feel sorry for *A*, if they are not playing the game of Figure 1?

(2) The (refined) consequence is *logically impossible* outside the specific game; for example, imagine we redefined *g* as "getting $3 after player *A* has chosen Up": clearly this is inconsistent with the game that *B* is playing in, say, Figure 2.

To sum up: when too much is included in the description of outcomes, the consequence itself remains tied to the specific game and cannot be used to construct arbitrarily other acts (or functions from states of the world to consequences).[39] Refining options to include reciprocity considerations clearly raises problems of both kinds. If players care about "what might have been," then this must be included in the definition of a consequence. Indeed the consequences must be refined so as to include entire branches of the game. But then one cannot use such consequences to create arbitrarily other acts, if such acts are to make sense at all. The Savage measurement procedure is not flexible enough to neutralize reciprocity counterexamples.[40]

---

[39] See also Sugden, "Rational Choice"; Munier, "Comment"; and Bruno Verbeek, "Consequentialism, Rationality, and the Relevant Description of Outcomes," *Economics and Philosophy*, XVII (2001): 181–205. These and related problems have been addressed in the literature by introducing so-called "state-dependent" utilities. At the price of introducing some extra axioms, it has been shown that individual choices can be described as maximizing a state-dependent utility function over acts. This leads to some problems, however, in the estimation of subjective probabilities; see, for example, Mark Schervish, Teddy Seidenfeld, and Joseph Kadane, "State-Dependent Utilities," *Journal of the American Statistical Association*, LXXXV (1990): 840–47. Such problems can be overcome if one is ready to abandon the Savage measurement procedure and especially the strong operationalism associated with it; see, for example, Edi Karni and Philippe Mongin, "On the Determination of Subjective Probability by Choices," *Management Science*, XLVI (2000): 233–48.

EVIDENCE OF RECIPROCITY

Reciprocity has been discussed by experimental game theorists for more than two decades now. Most of this discussion however has taken place *before* the existence of reciprocal motives had been established empirically. Consider social dilemma games again. As already mentioned, many subjects seem willing to cooperate in a one-shot sequential Prisoner's Dilemma game. But this behavior of course does not constitute decisive evidence in favor of reciprocity, because the mutual cooperation outcome provides an *equal* distribution of the material payoffs (see Figure 1 above) and therefore might be preferred by players that are averse to inequality.

A crucial test for reciprocity must implement some variant of the Savage measurement procedure. Players' preferences over a set of outcomes must be observed *first* "in isolation," and *then* in the context of a game that is likely to trigger reciprocity considerations (or more generally a concern for "what might have been"). The comparison between choices in these two contexts will provide direct evidence for reciprocity as well as a proper test of game theory. Several experimenters have done precisely this in recent years, although experimental game theorists themselves do not seem to have noticed the general implications of these results.

The experiments usually start with so-called Dictator's games. A Dictator's game[41] is not even a strategic game, strictly speaking, for it involves a straight choice between different allocations. The outcome is not determined by an interaction between the players: one player chooses, and the other takes whatever has been chosen. A Dictator's game then, like a stripped-down version of the Savage measurement procedure, simply elicits the (ordinal) structure of preferences of the experimental subjects. Since the ordering, rather than the cardi-

---

[40] Regret is another motivation that can lead to problems with the rectangular field assumption not only in the realm of game theory, but also in individual (nonstrategic) decision making. See, for example, Sugden, "Regret, Recrimination and Rationality," *Theory and Decision*, XIX (1986): 77–99. Unlike reciprocity, however, we do not have strong evidence about regret at the moment.

[41] See, for example, Robert Forsythe, Joel Horowitz, N.E. Savin, and Martin Sefton, "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior*, VI (1994): 347–69.

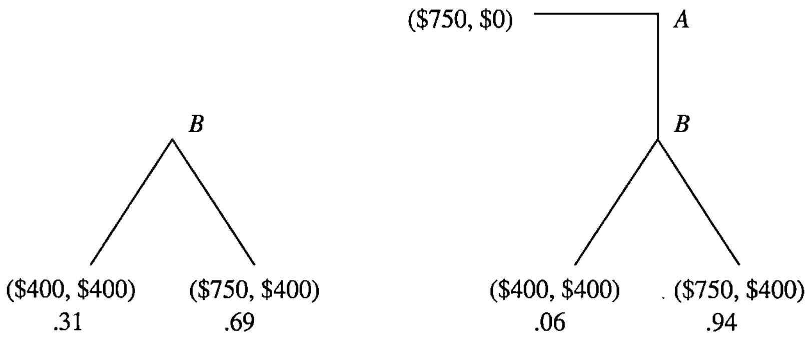


Figure 3

nality, of preferences is what matters in many simple experimental games, the full-blown Savage method is unnecessary.

Once preferences have been elicited in the Dictator's game, it is then possible to check whether players behave rationally in more complex strategic situations. Take the sequential Prisoner's Dilemma. Cox (*op. cit.*) uses a slightly more complex version of this setting, called the "Investment game." Player *A* here chooses how much of a given endowment to pass over to *B*. Whatever *A* chooses, the sum will be tripled by the experimenter. Player *B* then decides how much of this investment to return to *A*.

After the game is played, Cox offers some subjects exactly the same sums which have been selected by players of type *A* in the first experiment—except that type-*B* players this time do not know that it is an investment from another player (the sum is presented as a "pie from the sky"). The experimenter can therefore check whether the behavior of *B* players in this Dictator's game is radically different from their behavior in the Investment game. As a matter of fact it is: in the Investment task the average amount returned by *B* players is more than twice the amount allocated to *A* in the Dictator's task ($4.94 versus $2.06).

Gary Charness and Matthew Rabin[42] use even simpler settings (like, for example, Figure 3). The first part of the experiment (on the left) is again a Dictator's game. The first payoff between brackets is *A*'s monetary gain and the second one is *B*'s, as usual. The figures under each outcome represent the observed distribution of player *B*'s

---

[42] "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, CXVII (2002): 817–69.

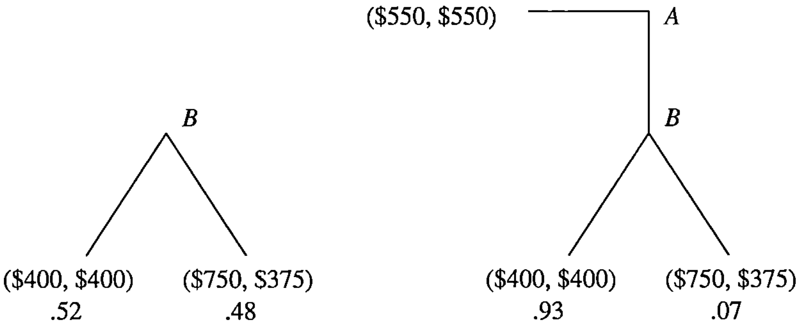


Figure 4

choices. A majority of subjects goes for the Pareto-superior outcome, but there is significant evidence of inequality aversion too.

The same (sub-)game is then transferred in a different context, where *B* is asked to make the same choice *after* player *A* has already made a move (Figure 3, on the right). As we can see, choices in this second setting differ substantially from those in the simple Dictator's game. Concerns for equality in this specific case seem to be largely swept away by *B*'s desire to respond to *A*'s kindness (*A* could have chosen unilaterally to give *B* absolutely nothing, after all).

Figure 4 presents another example of the importance of "what might have been." In the Dictator's game (left), *B* players display a considerable willingness to sacrifice a small part of their material payoff to increase *A*'s income. This willingness disappears in the enhanced game (right), where *B* moves knowing that *A* has refused the (550, 550) split (a phenomenon that Charness and Rabin dub "concern withdrawal").

There are many more experiments confirming the importance of reciprocity.[43] Much current work in experimental game theory is now focusing on the robustness of these patterns.[44] It is already possible to say, however, that in a number of settings *the structure of the game matters.* Subjects do not just care about the material consequences but also about the alternative moves that are available and the process that leads to each consequence. They care about what might have been and how they got where they are now.

---

[43] See, for example, Armin Falk, Ernst Fehr, and Urs Fischbacher, "On the Nature of Fair Behavior," *Economic Inquiry*, XLI (2003): 20–26; Kevin McCabe, Mary Rigdon, and Vernon Smith, "Positive Reciprocity and Intentions in Trust Games," *Journal of Economic Behavior and Organization*, LII (2003): 267–75.

[44] Reciprocity seems to be a fairly erratic phenomenon. Cristina Bicchieri argues, convincingly in my view, that social norms (such as reciprocity) are highly context-sensitive and usually must be "triggered" or "cued" by some factor signalling that "this is a context where the norm applies," that is, where it is generally expected that people behave according to the norm. See, for example, Bicchieri, *The Grammar of Society* (New York: Cambridge, 2006).

CONCLUDING REMARKS

Game theory is the piece of contemporary economics that has had the most dramatic impact outside the boundaries of the discipline. Sixty years from its inception, game theory is widely used in fields as different as political science, sociology, biology, psychology, law, military strategy, and, last but not least, philosophy. Its empirical status, then, is an issue of wide scientific and philosophical relevance.

For many years a considerable degree of confusion about the status of game theory was fostered by the proliferation of seemingly anomalous but indecisive evidence. Now that experiments of the right kind have been done—experiments that really test game theory—it has become impossible to argue that the anomalies can be digested as usual. In order to see that, it is necessary to appreciate the crucial distinction between the theory of rational play and the bridge principles that connect the former with measurable properties of the real world. These two elements are sometimes conflated, sometimes treated as if they were completely independent. This probably leads to the (mistaken) belief that game theory is as flexible an explanatory tool as the folk-psychological theory of action of which it constitutes a formal refinement.

The Savage procedure (and the rectangular field assumption upon which it is based) imposes some limits on the amount of information that can be incorporated in the description of the outcomes of a game. These technical limitations make sure that the spirit of genuine consequentialism (players are forward looking, subscribe to sub-game perfect rationality, and so on) is preserved.

It is important to understand the precise implications of what was just said: nothing prevents us from applying the tools of game and utility theory across a limited range of experimental games, assuming the subjects have consistent preferences over a limited number of highly specific (or highly refined) options, once the outcomes have been (re-)described so as to include the overall structure of the game.[45] Such preference information might turn out to be useless outside the specific game situation, of course, but could still be used to predict and summarize choices in a limited set of games. What cannot be done, if consequences become overly sophisticated, is to have a unique utility function across the set of *all* the games that can be constructed using those (and other) consequences.

---

[45] The Savage procedure prescribes *first* to measure the utility of the outcomes of a game in isolation and *then* to see how the subjects behave in more complex strategic situations (games). The rectangular field assumption is required if one wants to elicit utility measures for a *wide* range of consequences, or in other words if one is trying to specify the *fine-grained* structure of an individual's utility function. But the application of game theory in specific instances of course does not require this full-blown measurement to be carried out in practice. In fact, as we have seen, the simple games that are customarily tested in the laboratory often require only a simplified or stripped-down version of the Savage procedure—because the mere ordinal ranking of a handful of outcomes is what really matters.

From this respect, the "naive" and often ridiculed convention of using monetary payoffs as initial conditions or proxies of players' utilities, has much to recommend. It is well known among historians and philosophers of science that the success of scientific theories is often crucially dependent on the creation of instruments of observation that allow adequate measurements of their key variables. Consider Galileo's telescope, but also the microscope for the atomistic research program, the steam engine for thermodynamics, and so forth. Monetary payoffs are objective and observable, and provided we have the right theory at hands, can be used readily to generalize to other situations that are similar (in the monetary payoffs structure) to those that were tested in the laboratory. If this route is followed of course the "core" of game theory must be modified to account for the existing evidence. The partnership between the standard theory of rational play and the Savage procedure has proven to be unsustainable (given the evidence). We can keep the core of game theory and reject the bridge principles (the Savage procedure), but face difficult problems of applicability. The other viable alternative is to keep the bridge and change the theory of play.

Why did it take such a long time to design and run the right experiments to test game theory? One reason is that, surprisingly, the Savage procedure was never part of the official toolkit of the experimental game theorist. This is the result of complicated historical factors, in particular the fact that the methodology of experimental economics (especially so-called "induced value theory"[46]) was developed mainly in the context of market experiments and was aimed at the active *inducement*, rather than the passive observation of preferences. But a proper discussion of these issues would require a long detour into the history and methodology of experimental game theory, which cannot be attempted here.[47]

Interestingly, the refutation of game theory took place almost simultaneously with the theory's most remarkable successes in policy-making. The reform of inefficient (usually centralized) methods of allocation and the design of new markets (for electricity, telecommunication, health services, and so on) is now increasingly informed by the principles of game theory. These principles moreover are often subjected to preliminary testing in controlled conditions before being applied to real-world situations. These tests, however, are usually aimed at spotting and avoiding precisely the sort of problematic situations that the theory is unable to deal with. Other-regarding preferences for instance can simply be ruled out by design, by making sure that the available information and the structure of the game prevent such concerns from arising in the first place. This is economic engineering, as opposed to theory-testing: the world is modified so as to "fit" the conditions that allow the theory to work. But since engineering successes often *do* drive science, we should not rush to any conclusion about the future of game theory. Its refutation and its remarkable success can (and will) coexist in the minds of many social scientists for whom the theory is, and is likely to remain, the best game in town.

---

[46] See, for example, Vernon Smith, "Microeconomic Systems as an Experimental Science," *American Economic Review*, LXXII (1982): 923–55.

[47] See, however, Guala, *The Methodology of Experimental Economics* (New York: Cambridge, 2005), especially chapter 11.

FRANCESCO GUALA

University of Exeter

NEW BOOKS: REPRINTS

FLEW, ANTONY: *God and Philosophy*. Amherst, NY: Prometheus, 1966/1984/2005. 210 p. Paper $18.00.

FORSTER, MICHAEL N.: *Wittgenstein on the Arbitrariness of Grammar*. Princeton: University Press, 2004/2005. xiii, 247 p. Cloth $45.00, paper $16.95.

HACKING, IAN: *Historical Ontology*. Cambridge: Harvard, 2002/2004. vii, 279 p. Cloth $45.00, paper $17.95.

JOUBERT, JOSEPH: *The Notebooks of Joseph Joubert*. PAUL AUSTER, trans. New York: Review of Books, 1983/2005. xv, 159 p. Paper $14.95.

KEARNEY, RICHARD, ed.: *Twentieth-Century Continental Philosophy: Routledge History of Philosophy, Volume VIII*. New York: Routledge, 1999/2004. xlv, 524 p. Paper $32.95.

KENNY, ANTHONY: *Wittgenstein*. Malden, MA: Blackwell, 1973/2006. xx, 191 p. Cloth $74.95, paper $29.95.

LEVINE, MICHAEL P.: *The Analytic Freud: Philosophy and Psychoanalysis*. New York: Routledge, 2000/2005. xii, 320 p. Cloth $135.00, paper $36.95.

LUGG, ANDREW: *Wittgenstein's "Investigations" 1–133: A Guide and Interpretation*. New York: Routledge, 2000/2004. xi, 213 p. Cloth $85.00, paper $36.29.

MACINTYRE, ALASDAIR: *The Unconscious: A Conceptual Analysis, Revised Edition*. New York: Routledge, 1958/2004. v, 122 p. Cloth $60.00, paper $20.99.

MARTINICH, A.P.: *Philosophical Writing: An Introduction, 3rd edition*. Malden, MA: Blackwell, 1996/2005. 202 p. Paper $19.95.

MCCABE, HERBERT: *God Matters*. New York: Continuum, 1987/2005. 249 p. Paper $19.95.

MCGUINNESS, BRIAN: *Young Ludwig: Wittgenstein's Life, 1889–1921*. New York: Oxford, 1988/2005. xiv, 322 p. Paper $29.95.

MEYER, JAMES: *Minimalism: Art and Polemics in the Sixties*. New Haven: Yale, 2001/2004. viii, 340 p. Cloth $75.00, paper $35.00.

MILL, JOHN STUART: *On Liberty*. STEVEN M. CAHN and C.L. TEN, eds. Lanham, MD: Rowman and Littlefield, 2005. 180 p. Paper $9.95.

MILLER, MITCHELL H., JR.: *The Philosopher in Plato's Statesman*. Las Vegas: Parmenides, 1980/2004. xx, 186 p. Paper $20.00.

MISAK, C.J.: *Truth and the End of Inquiry: A Peircean Account of Truth*. New York: Oxford, 1991/2004. Cloth $90.00, paper $29.95.

MOHR, RICHARD D: *God and Forms in Plato*. Las Vegas: Parmenides, 1985/2005. xxv, 279 p. Paper $28.00.

MOORE, G.E.: *Ethics*. New York: Oxford, 1966/2005. xliii, 174 p. Cloth $74.00, paper $27.50.

NAGEL, THOMAS: *Concealment and Exposure*. New York: Oxford, 2002/2004. 240 p. Cloth $39.95, paper $18.95.

ONG, WALTER J.: *Ramus, Method, and the Decay of Dialogue*. Chicago: University Press, 1975/2005. xxvii, 408 p. Paper $22.50.

PUTNAM, HILARY: *Ethics without Ontology*. Cambridge: Harvard, 2004/2005. ix, 160 p. Cloth $29.95, paper $16.95.

RAWLS, JOHN: *Political Liberalism: Expanded Edition*. New York: Columbia, 1993/1996/2005. lx, 525 p. Cloth $64.50, paper $24.50.

REGAN, TOM: *The Case for Animal Rights, Updated with a New Preface.* Berkeley: California UP, 1983/1985/2004. lv, 425 p. Paper $21.95.

ROSS, GEORGE MACDONALD, and TONY MCWALTER, eds.: *Kant and his Influence.* New York: Continuum, 1990/2005. xv, 374 p. Cloth $60.00, paper $39.95.

RUSE, MICHAEL: *Darwin and Design: Does Evolution Have a Purpose?* Cambridge: Harvard, 2003/2004. x, 371 p. Cloth $29.95, paper $16.95.

SARTRE, JEAN-PAUL: *Critique of Dialectical Reason, Volume One: Theory of Practical Ensembles.* ALAN SHERIDAN-SMITH, trans., JONATHAN RÉE, ed. New York: Verso, 1976/1991/2004. xxxiii, 836 p. Paper $25.00.

SCRUTON, ROGER: *Philosophy: Principles and Problems.* New York: Continuum, 1996/2005. 168 p. Paper $17.95.

SHANKER, STUART G., ed.: *Philosophy of Science, Logic and Mathematics in the Twentieth Century: Routledge History of Philosophy, Volume IX.* New York: Routledge, 1996/2003. xxxviii, 461 p. Cloth $195.00, paper $32.95.

SUITS, BERNARD: *The Grasshopper: Games, Life and Utopia.* Orchard Park, NY: Broadview, 1978/2005. 179 p. Paper $22.95.

SWINBURNE, RICHARD: *The Existence of God, Revised Edition.* New York: Oxford, 1979/1984/1991/2004. vi, 363 p. Cloth $95.00, paper $24.95; *Faith and Reason.* New York: Oxford, 1981/2005. vi, 278 p. Cloth $99.00, paper $35.00.

WHITE, MORTON: *A Philosophy of Culture: The Scope of Holistic Pragmatism.* Princeton: University Press, 2002/2005. xv, 193 p. Paper $18.95.

WOOD, ALLEN W.: *Karl Marx, Second Edition.* New York: Routledge, 1981/2004. xli, 302 p. Cloth $104.95, paper $34.95.

YOUNG-BRUEHL, ELISABETH: *Hannah Arendt: For Love of the World, Second Edition.* New Haven: Yale, 1982/2004. liii, 563 p. Paper $22.00.

NEW BOOKS: TRANSLATIONS

ALTHUSSER, LOUIS: *Philosophy of the Encounter: Later Writings, 1978–1987.* G.M. GOSHGARIAN, trans. New York: Verso, 2006. l, 299 p. Cloth $90.00, paper $27.00.

AMBROSE OF MILAN: *Ambrose of Milan: Political Letters and Speeches.* WOLFE LIEBESCHUETZ, trans. Liverpool: University Press, 2005. x, 424 p. Cloth $85.00.

AQUINAS, THOMAS: *Aquinas: Disputed Questions on the Virtues.* E.M. ATKINS, trans. New York: Cambridge, 2006. xl, 301 p. Cloth $75.00; *Summa Theologiae, Questions on God.* BRIAN DAVIES, trans. New York: Cambridge, 2006. xli, 298 p. Cloth $75.00, paper $29.99.

ARISTOTLE: *Physics or Natural Hearing.* GLEN COUGHLIN, trans. South Bend, IN: St. Augustine's, 2005. xxxi, 308 p. Cloth $40.00, paper $18.00.

ASVAGHOSHA: *The Awakening of Faith: Attributed to Asvaghosha.* YOSHITO S. HAKEDA, trans. New York: Columbia, 2006. xii, 136 p. Cloth $49.50, paper $22.50.

BADIOU, ALAIN: *Metapolitics.* JASON BARKER, trans. New York: Verso, 2005. xxxviii, 159 p. Cloth $27.00; *Being and Event.* OLIVER FELTHAM, trans. New York: Continuum, 2005. xxxiii, 526 p. Cloth $29.95.

BARBARAS, RENAUD: *Desire and Distance: Introduction to a Phenomenology of Perception.* PAUL B. MILAN, trans. Stanford: University Press, 2006. ix, 169 p. Cloth 55.00, paper $22.95.

CASATI, ROBERTO, and ACHILLE VARZI: *Insurmountable Simplicities: Thirty-nine Philosophical Conundrums.* ACHILLE VARZI, trans. New York: Columbia, 2006. 132 p. Cloth $24.95.

CHERKI, ALICE: *Franz Fanon: A Portrait.* NADIA BENABID, trans. Ithaca, NY: Cornell, 2006. x, 255 p. Cloth $59.95, paper $24.95.

CIXOUS, HÉLÈNE: *Dream I Tell You.* BEVERLEY BIE BRAHIC, trans. New York: Columbia, 2006. 160 p. Cloth $29.50.

FICHTE, J.G.: *The System of Ethics.* DANIEL BREAZEALE and GÜNTER ZÖLLER, trans. New York: Cambridge, 2005. xliv, 399 p. Cloth $75.00, paper $29.99.

HÖFFE, OTFRIED: *Kant's Cosmopolitan Theory of Law and Peace.* ALEXANDRA NEWTON, trans. New York: Cambridge, 2005. xvii, 253 p. Cloth $70.00, paper $24.99.

JAINCAUD, DOMINIQUE: *On the Human Condition.* EILEEN BRENNAN, trans. New York: Routledge, 2005. xxiv, 71 p. Paper $15.95.

KANT, IMMANUEL: *Kant: Anthropology from a Pragmatic Point of View.* ROBERT B. LOUDEN, trans. New York: Cambridge, 2006. xxxix, 246 p. Cloth $65.00, paper $22.99.

KEUTH, HERBERT: *The Philosophy of Karl Popper.* HERBERT KEUTH, trans. New York: Cambridge, 2004. xiv, 367 p. Cloth $75.00, paper $29.99.

LEFEBVRE, HENRI: *Critique of Everyday Life, Volume III: From Modernity to Modernism.* MICHEL TREBITSCH, trans. New York: Verso, 2006. xxxiv, 179 p. Cloth $35.00.

LEIBNIZ, G.W.: *Confessio Philosophi: Papers Concerning the Problem of Evil, 1671–1678.* ROBERT C. SLEIGH, JR., trans. New Haven: Yale, 2005. xli, 178 p. Cloth $85.00.

LESSING, GOTTHOLD EPHRAIM: *Lessing: Philosophical and Theological Writings.* H.B. NISBET, trans. New York: Cambridge, 2005. xv, 262 p. Cloth $70.00, paper $27.99.

MARCEL, GABRIEL: *Music and Philosophy.* STEPHEN MADDUX and ROBERT E. WOOD, trans. Milwaukee, WI: Marquette, 2005. 147 p. Paper $17.00.

NIETZSCHE, FRIEDRICH: *Prefaces to Unwritten Works.* MICHAEL W. GRENKE, trans. South Bend, IN: St. Augustine's, 2005. viii, 141 p. Cloth $18.00.

PLATO: *Lysis.* TERRY PENNER and CHRISTOPHER ROWE, trans. New York: Cambridge, 2005. xiv, 366 p. Cloth $95.00.

QUINTILIAN: *The Lesser Declamations, Volume I and II.* D.R. SHACKLETON BAILEY, trans. Cambridge: Harvard, 2006. vii, 445 p. viii, 467 p. Cloth $21.50.

REINHOLD, KARL LEONHARD: *Letters on the Kantian Philosophy.* JAMES HEBBELER, trans. New York: Cambridge, 2005. l, 230 p. Cloth $70.00, paper $27.99.

RICOEUR, PAUL: *The Course of Recognition.* DAVID PELLAUER, trans. Cambridge: Harvard, 2005. xi, 297 p. Cloth $29.95.

ROUSSEAU, JEAN-JACQUES: *The Plan for Perpetual Peace, On the Government of Poland, and Other Writings on History and Politics.* CHRISTOPHER KELLY and JUDITH BUSH, trans. Hanover, NH: Dartmouth, 2005. xxiii, 260 p. Cloth $70.00.

SALA-MOLINS, LOUIS: *Dark Side of the Light: Slavery and the French Enlightenment.* JOHN CONTEH-MORGAN, trans. Minneapolis: Minnesota UP, 2006. xxxvi, 165 p. Cloth $56.95, paper $18.95.

THE JOURNAL OF PHILOSOPHY

BACK ISSUE SALE

Complete volumes
are now available
at the following discounted prices
in order to reduce our inventory:

Volumes 1 through 98	$10 each
Volumes 99 through 101	$25 each

For more information, contact Pamela Ward,
Business Manager, at pcw1@columbia.edu

Visa, Mastercard, American Express and
PayPal are accepted through our website at
www.journalofphilosophy.org

New from Stanford

H. C. for Life, That Is to Say...

JACQUES DERRIDA
Translated by Laurent Milesi and Stefan Herbrechter

H. C. for Life, That Is to Say... is Jacques Derrida's tribute to Hélène Cixous—the author, her works, and their lifelong mutual reading and intellectual friendship.

Meridian: Crossing Aesthetics
$21.95 paper $55.00 cloth

AVAILABLE IN JULY 2006

Prostitution and Pornography
Philosophical Debate About the Sex Industry

Edited by JESSICA SPECTOR

"Jessica Spector has compiled the most challenging and thorough volume yet on philosophic themes surrounding the vexed questions of pornography and prostitution. . . . This will be a seminal work in the field."
—Drew A. Hyland, Trinity College

$27.95 paper $65.00 cloth

The End of Art
Readings in a Rumor after Hegel

EVA GEULEN
Translated by James McFarland

Readings of Hegel, Nietzsche, Benjamin, Adorno, and Heidegger trace the role that the discourse on the end of art has played in post-Hegelian philosophical aesthetics.

$19.95 paper $50.00 cloth

Reflections of Equality

CHRISTOPH MENKE
Translated by Howard Rouse and Andrei Denejkine

This book argues that the center of political modernity is determined by a conflictive relation between the liberal core concept of political equality and the idea of individuality.

Cultural Memory in the Present
$24.95 paper $65.00 cloth

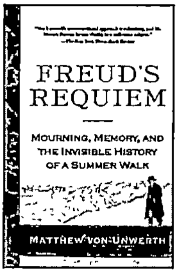
Stanford
University Press
800.621.2736 www.sup.org

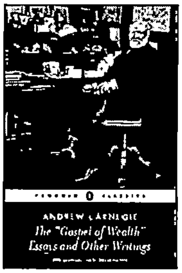
You can now subscribe to the *Journal of Philosophy* using

PayPal

To learn more about this easy new option, click on the PayPal icon at **www.jphil.org.**

NEW IN PHILOSOPHY FROM
PENGUIN GROUP (USA)


FREUD'S REQUIEM
MOURNING, MEMORY, AND THE INVISIBLE HISTORY OF A SUMMER WALK
MATTHEW VON UNWERTH



PLATO
PROTAGORAS AND MENO
Translated by Adam Beresford
Introduction by Lesley Brown
Penguin Classics 176 pp. 0-14-044903-5 $10.00

DEBORAH BLUM
GHOST HUNTERS
William James and
the Search for Scientific
Proof of Life after Death
The Penguin Press 368 pp. 1-59420-090-4 $25.95

MATTHEW VON UNWERTH
FREUD'S REQUIEM
Mourning, Memory, and the
Invisible History of a Summer Walk
Riverhead 256 pp. 1-59448-199-7 $15.00

JOSEPH MAZUR
EUCLID IN THE RAINFOREST
Discovering Universal
Truth in Logic and Math
Plume 352 pp. 0-452-28783-9 $15.00

BART KOSKO
NOISE
A revelatory look at the phenomenon of noise from the man who introduced the concept of fuzzy logic.
Viking 320 pp. 0-670-03495-9 $24.95

XENOPHON
HIERO THE TYRANT AND OTHER TREATISES
Translated by Robin A. H. Waterfield
Introduction and Notes by Paul Cartledge
Penguin Classics 304 pp. 0-14-045525-6 $16.00

PLUTARCH
THE FALL OF THE ROMAN REPUBLIC
Translated by Rex Warner
Edited with an Introduction
and Notes by Robin Seager
Penguin Classics 464 pp 0-14-044934-5 $15.00

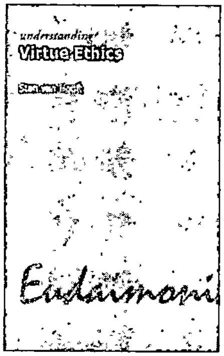
ANDREW CARNEGIE
THE "GOSPEL OF WEALTH" ESSAYS AND OTHER WRITINGS
Edited with an Introduction
and Notes by David Nasaw
Penguin Classics 400 pp. 0-14-303989-X $10.00

JULIAN BAGGINI
THE PIG THAT WANTS TO BE EATEN
100 Experiments for
the Armchair Philosopher
Plume 336 pp. 0-452-28744-8 $14.00

For a free **PHILOSOPHY** brochure please email: academic@penguin.com

PENGUIN GROUP (USA) www.penguin.com/academic
Academic Marketing Department, 375 Hudson St., NY, NY 10014

A New Series

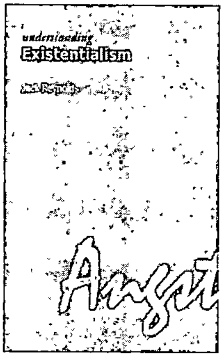


Understanding Virtue Ethics
Stan van Hooft

Acumen Publishing
Understanding Movements in Modern Thought

1-84465-045-6	$22.95	paper
1-84465-044-8	$85.00	cloth

"Clear and concise with imaginative and well-drawn examples. The book displays an excellent grasp of the main themes in virtue ethics." Michael Brady, University of Glasgow

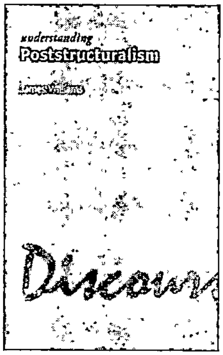


Understanding Existentialism
Jack Reynolds

Acumen Publishing
Understanding Movements in Modern Thought

1-84465-043-X	$22.95	paper
1-84465-042-1	$85.00	cloth

Understanding Existentialism begins with a discussion of Kierkegaard, Nietzsche, Husserl, Jaspers, and Marcel – existentialism's antecedents – and ends with an assessment of the movement's effect, particularly its influence on poststructuralism.



Understanding Poststructuralism
James Williams

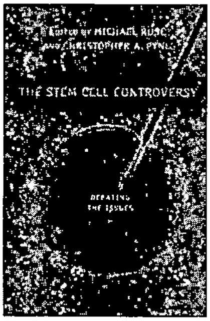
Acumen Publishing
Understanding Movements in Modern Thought

1-84465-033-2	$22.95	paper
1-84465-032-4	$85.00	cloth

"Williams establishes himself as one of the foremost English-language interpreters of French thought and presents a judicious evaluation of the poststructuralist movement as a whole."
Daniel W. Smith, Purdue University

McGill
Queen's
University
Press

www.mqup.ca



THE STEM CELL CONTROVERSY

Debating the Issues, Second Edition

Edited by Michael Ruse and Christopher A. Pynes

350 pp • ISBN 1-59102-404-8 • PB • $22 • April 2006

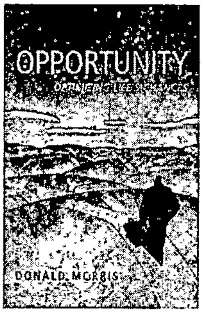
The contributors—scientists, medical practitioners, philosophers, theologians, historians, and policy analysts—offer a variety of perspectives to give readers the critical tools they need to shape an informed position on the topic. Readers will come away with a deeper understanding of the science of stem cell research, its medical cures and promises, and the moral, religious, and policy concerns surrounding this controversial social issue.

OPPORTUNITY

Optimizing Life's Chances

Donald Morris

390 pp • ISBN 1-59102-402-1 • HC • $27 • April 2006



"Donald Morris's Opportunity...considers choices that promise a significant change in the life of the person making the choice.... Morris's survey of the literature is wide-ranging in terms of both literary sources and topics considered, thus offering insights to a variety of readers." **—Elizabeth R. Eames**
Professor Emerita, Philosophy Department
Southern Illinois University

UNCORRECTED PAPERS

Diverse Philosophical Dissents

Wallace Matson

350 pp • ISBN 1-59102-397-1 • HC • $35

These incisive, witty, and completely accessible essays on a wide range of topics by historian of philosophy Wallace Matson admirably demonstrate that philosophy can still be based on careful reasoning and presented with clarity of expression. Against fashionable contemporary views, Matson asserts that philosophy is "the most important subject in the college curriculum," because it is the investigation into what rationality is. Getting the answer wrong to the question "What does it mean to be reasonable?" is the most catastrophic of errors. The motivation for most of the essays in this collection is Matson's perception that this error is being widely committed and that received opinion on many topics is dead wrong.

THE HEAD AND THE HEART

Philosophy in Literature

Burton F. Porter

170 pp • ISBN 1-59102-411-0 • HC • $30 • May 2006

Part of the greatness of great literature consists in the profound, philosophic ideas the works contain. These ideas may not be unknown to philosophy but, when rendered in literary form, they gain an aesthetic force often lacking in the philosophic treatise with its careful train of reasoning.

In this insightful study, Burton Porter explores the philosophic content of some outstanding literary works, analyzing and evaluating the ideas that drive the narrative.

Prometheus Books
Toll Free: 800-421-0351 • Fax: 716-691-0137 • www.prometheusbooks.com

Philosophy from YALE

A Dictionary of Continental Philosophy

Edited by JOHN PROTEVI

With over 450 definitions and articles, this comprehensive dictionary encompasses all the important thinkers, topics, and technical terms associated with continental philosophy.

"Everywhere I turned I found **elegant summaries.** There is an extraordinary array of briefs on important figures and themes and dozens of explanations of technical concepts, many notorious for their obscurity."—Wayne M. Martin

"This dictionary is marked by an unusual comprehensiveness that makes it of **considerable value to specialists and students alike.**"
—Stanley Rosen $75 00

Manliness

HARVEY C. MANSFIELD

Covering subjects from Hemingway to Plato, this provocative book presents the first comprehensive study of manliness—what it is and whether it's obsolete.

"A **thoughtful** attempt to move us to think more clearly about who we are, and about the future of our liberal society."
—Mary Nichols $27.50

Distributive Justice and Disability

Utilitarianism against Egalitarianism

MARK S. STEIN

"An **important, lucid, powerfully argued** book that sheds fresh light on the debate between utilitarians and egalitarians."—Peter Singer

"A thoughtful and impressive book that makes an important contribution to utilitarian moral philosophy."—Julia Driver
$50.00

New in Paper

Politics and Passion

Toward a More Egalitarian Liberalism

MICHAEL WALZER

"This clearly and **engagingly written** book shows up the thinness of much contemporary Liberal theory."—Charles Taylor

"An elegant and probing critique of contemporary liberal thought."
—G. John Ikenberry, *Foreign Affairs*
$15.00 paperback

The Jewish Political Tradition

Volume Two: Membership

Edited by MICHAEL WALZER, MENACHEM LORBERBAUM, and NOAM J. ZOHAR

"Every one of the chapters **enacts a debate that should have a living resonance,** not just for Jews, although obviously for them, but for everyone with a historical sense and a political conscience."
—Hilary Putnam, *Boston Review*
$30.00 paperback

YALE University Press • yalebooks.com

New from
Oxford

Two-Dimensional Semantics

EDITED BY MANUEL GARCIA-CARPINTERO AND JOSEP MACIÀ

This selection of new essays by some of the world's leading authorities in this field sheds fresh light both on foundational issues regarding two-dimensional semantics and on its specific applications.

2006 paper $35 00 cloth $99 00

Metaphysical Essays

JOHN HAWTHORNE

John Hawthorne is widely regarded as one of the finest philosophers working today. He is perhaps best known for his contributions to metaphysics, and this volume collects his most notable papers in this field.

2006 paper $35.00 cloth $99.00

In Contradiction

Second Edition

GRAHAM PRIEST

In Contradiction advocates and defends the view that there are true contradictions (dialetheism), a view that flies in the face of orthodoxy in Western philosophy since Aristotle. This second edition of the book substantially expands upon the original in various ways, and also contains the author's reflections on developments over the last two decades.

2006 paper $35.00 cloth $99.00

Ignorance and Imagination

The Epistemic Origin of the Problem of Consciousness

DANIEL STOLJAR

Ignorance and Imagination advances a novel way to resolve the central philosophical problem about the mind: how it is that consciousness or experience fits into a larger naturalistic picture of the world.

(Philosophy of Mind Series)
2006 $49 95

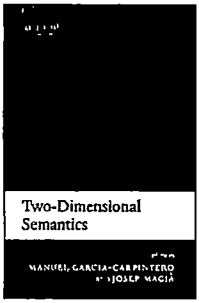
Normativity and the Will

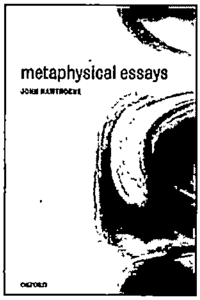
Selected Essays on Moral Psychology and Practical Reason

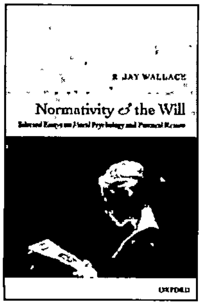
R. JAY WALLACE

Normativity and the Will collects fourteen important papers on moral psychology and practical reason by R. Jay Wallace, one of the leading philosophers currently working in these areas.

2006 paper $35.00 cloth $95.00







Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS

CAMBRIDGE

BEST IN SCHOLARSHIP

The Cambridge Companion to Keynes
Edited by Roger Backhouse
and Bradley Bateman
Cambridge Companions to Philosophy
$70.00: Hb: 0-521-84090-2: 342pp
$28.99: Pb: 0-521-60060-X

Conventionalism
From Poincare to Quine
Yemima Ben-Menahem
$80.00: Hb: 0-521-82619-5: 340pp

Understanding Space-Time
The Philosophical
Development of Physics
from Newton to Einstein
Robert DiSalle
$75.00: Hb: 0-521-85790-2: 188pp

The Problem of Animal Generation in Early Modern Philosophy
Edited by Justin Smith
Cambridge Studies in Philosophy and Biology
$85.00: Hb: 0-521-84077-5: 470pp

The Tasks of Philosophy
Selected Essays
Volume 1
Alasdair MacIntyre
$70.00: Hb: 0-0521-85437-7: 248pp
$24.99: Pb: 0-521-67061-6

Ethics and Politics
Selected Essays
Volume 2
Alasdair MacIntyre
$70.00: Hb: 0-521-85438-5: 256pp
$24.99: Pb: 0-521-67062-4

Knowledge and Inquiry
Essays on the
Pragmatism of Issac Levi
Erik J. Olsson
Cambridge Studies in Probability, Induction and Decision Theory
$80.00: Hb: 0-521-84556-4: 400pp

Moral Disagreement
Folke Tersman
$60.00: Hb: 0-521-85338-9: 160pp

**prices subject to change*

*For more information, please visit us
at www.cambridge.org/us
or call toll-free at* 1-800-872-7423

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

FROM CAMBRIDGE

The New Hegelians
Politics and Philosophy in the Hegelian School
Douglas Moggach
$75.00: Hb: 0-521-85497-0: 360pp

Extending Mechanics to Minds
The Mechanical Foundations of Psychology and Economics
Jon Doyle
$95.00: Hb: 0-521-86197-7: 480pp

Philosophical Perspectives on Infinity
Graham Oppy
$75.00: Hb: 0-521-86067-9: 334pp

Rethinking the Ontological Argument
A Neoclassical Theistic Response
Daniel A. Dombrowski
70.00: Hb: 0-521-86369-4: 176pp

Habermas and Theology
Nicholas Adams
$75.00: Hb: 0-521-86266-3: 278pp
$29.99: Pb: 0-521-68114-6

And announcing...

Cambridge Companions Online

- **The Literature and Classics Companions Collection**
- **The Philosophy, Religion and Culture Companions Collection**
- **The Complete Companions Collection**

To sign up or for a FREE TRIAL today, please write us at
ccol@cambridge.org

* *prices subject to change*

For more information, please visit us at www.cambridge.org/us
or call toll-free at 1-800-872-7423

CAMBRIDGE
UNIVERSITY PRESS

COMING SOON:

THE JOURNAL OF PHILOSOPHY

ONLINE

The Journal of Philosophy will be adding
a print and online subscription option for institutions
to begin Fall 2006

SUBSCRIPTION OPTIONS

Libraries and Institutions (print and online)	$180.00
Libraries and Institutions (print only)	$100.00
Individuals	$45.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

We expect to launch this new feature with our September 2006 issue.
Please consult our website at www.journalofphilosophy.org
for more details.

Patto
20/12/06

THE JOURNAL OF PHILOSOPHY

VOLUME CIII, NUMBER 6
JUNE 2006

page
267 *Absolutist Moral Theories and Uncertainty* — Frank Jackson and Michael Smith
284 *What Is Wrong with Weakness of Will?* — Alison McIntyre
BOOK REVIEWS
312 Jody Azzouni: *Deflating Existential Consequence: A Case for Nominalism* — Gideon Rosen

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Wolfgang Mann, Mary Mothersill, Christopher Peacocke, Carol Rovane, Achille C. Varzi. *Consulting Editors*: David Albert, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2006

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.
POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2006 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CIII, NO. 6, JUNE 2006

ABSOLUTIST MORAL THEORIES AND UNCERTAINTY*

You have promised to vote for Mary in the election for party president. You then learn that Fred's policies are much better. What to do? Most of us grant that if Fred's policies are a great deal better, you should break your promise. Absolutists about promise keeping demur. Bad consequences from keeping a promise never justify breaking it no matter how bad they may be. That is a very familiar issue. Our concern in this paper is with an issue that seems to have slipped under the radar. Suppose that you are pretty sure but far from certain that you promised to vote for Mary. It was some time ago; it was quite a party; you know that you indicated that you would vote for her, but did you actually *promise* to do so? What should absolutists about promise keeping say about how the element of doubt affects the question of what you ought to do in such a case?

We will be concerned with this question as it arises for absolutist theories in general. An absolutist theory absolutely prohibits actions of kind *K*, where *K* may be promise breaking, lying (as in Kant[1]), intentionally killing the innocent (as in some versions of traditional Catholic ethical theory), or else see the list of "things forbidden whatever consequences threaten" in the Hebrew-Christian ethic according to Elizabeth Anscombe.[2] Here we need to understand *K* as a property of an action as opposed to a relation between an action and available alternatives to that action. Classical utilitarianism absolutely prohibits doing actions that fail to maximize utility. But an action's failing to

---

*We are indebted to many audiences and the paper is substantially changed from the various presented versions. Special thanks to Andy Egan, Laura Schroeter, Karen Jones, Jim Griffin, and the editors of this JOURNAL.

[1] Kant, "On the Supposed Right to Lie from Philanthropic Motives," in *The Metaphysics of Morals* (*Die Metaphysik der Sitten*, 1797).

[2] Anscombe, "Modern Moral Philosophy," in *Ethics, Religion, and Politics: The Collected Papers of G.E.M. Anscombe*, Volume III (Minneapolis: Minnesota UP, 1981), pp. 26–42.

0022-362X/06/0306/267–83

maximize utility is a relation the action has to available alternatives. The distinctive feature of the kind of absolutism that we find, for example, in Kant and the Catholic tradition is that the absolutely prohibited kind is independent of the nature of any available alternatives. Of course, an important question for absolutisms of this kind is what to say about cases where *all* the available actions are of the prohibited kind. There are, say, only two things you can do and both involve breaking a promise (a different promise in each case). Opinions differ on how much of a problem this is for absolutism and our discussion will steer clear of the question except for a short mention later in the context of the bearing of strong dilemma views on our argument. Our examples will all be cases where at most one option is of the prohibited kind, *K*.[3]

More particularly, our concern will be with the following question: agent *S* has to choose between doing *A*, where *A* has p chance of being *K* and $1-p$ chance of being not-*K*, and doing *B*, where *B* is certain to have such-and-such consequences. What should a theory that absolutely prohibits doing *K* say concerning what *S* ought to do, as p rises and falls? We know what an absolutist theory says when $p = 1$. It says that no matter what the such-and-such consequences are, the action is forbidden. What about values of p other than 1?[4]

Mostly, we will frame the argument to come around theories that absolutely prohibit intentionally killing the innocent. They are more attractive and more popular than absolutisms that prohibit lying or promise breaking. However, it will be clear, we trust, that our discussion applies, *mutatis mutandis*, to absolutist theories in general. Our discussion will not be especially friendly to absolutism; the message will be that absolutism has serious trouble giving a plausible answer to our question. But before we enter the argument proper, we need to remark on a key distinction among "oughts."

I. TWO SENSES OF "OUGHT"

How should the epistemic state of an agent affect what they ought morally to do? In one sense of "ought," the answer is not at all; in another, the agent's epistemic state is crucial. We will make the point with a simple example.

---

[3] Thanks here to audiences and the editors.

[4] We could complicate the discussion by considering cases where there is doubt about the such-and-such consequences but we will not as the points we wish to make can be made without the complication.

A doctor has a patient with a minor but unpleasant skin condition on his leg. She knows that there is only one drug that is capable of curing the skin condition completely. She knows that in 95% of cases it does so without any deleterious side effects but that in 5% of cases it leads to the loss of the affected limb. There is no way of knowing which group any given patient falls into but, as it happens, our patient is one of the majority. What ought the doctor to do?

It is a mistake to insist that there is only one answer. There are two correct answers, each correct for a different sense of "ought." In one sense she ought not prescribe the drug. A 5% risk of limb loss is not justified by the potential benefit involved. If our doctor prescribed the drug, she would rightly be struck off the medical register. This is the sense of "ought" on which the agent's epistemic state matters. It is the sense of "ought" often called the subjective one. However, there is also a central sense of "ought" on which she ought to prescribe the drug. Suppose that the drug is not prescribed but that, some time in the future, a fail-safe test for whether a patient falls into the 95% or the 5% group is developed. It reveals that the patient in our example is in the 95% group. We will rightly say, looking back, something like "The right decision was made at the time, but we can now see that it would have been best to have prescribed the drug." But if it would have been best, there must be some good sense in which it ought to have been done. This is the sense in which what ought to be done has nothing to do with the agent's epistemic state, and is the sense of "ought" sometimes called the objective sense.

The subjective-objective terminology is unfortunate. It suggests that there is really one concept, namely, of *that which ought to be done,* and the distinction is between what *in fact* falls under that concept, which is what objectively ought to be done, and that which *probably* falls under that concept, which is what subjectively ought to be done.[5] However, what our doctor ought to do at the time of decision is not that which is most likely to be objectively right. It is 95% probable that prescribing the drug is the action which is objectively right, and that is precisely the course of action which she ought not do given what she knows at the time. There are two distinct concepts we need to inventory and when it is important to bear the distinction in mind, we will use "objective-ought" for the one according to which the drug in our example ought to have been prescribed, and "decision-ought" for the one on which the drug ought not to have been prescribed; that is, she objective-ought to have prescribed the drug and decision-ought not to have prescribed the drug. The "decision-ought" terminology has the drawback of possibly suggesting that we are talking about the right decision process to follow, whereas we are in fact talking about what a subject ought to do given their epistemic situation, but it has the signal advantage of capturing the key idea that it is the "ought" that matters most directly for what a subject ought to do at the time of acting and which determines the question of an agent's responsibility for an action.

---

[5] There is a distinction between that which is square in fact and that which is likely square but there are not two concepts *square.*

The example we used as a reminder of the objective-ought/decision-ought distinction is an example where what ought to be done is determined by consequentialist considerations. But it would be wrong to think of the distinction as an in-house issue for consequentialists that is of no interest to absolutists. To start with, almost every moral theory allows that sometimes what ought to be done is settled by the likely consequences, and whenever that is the case, it is transparent that the distinction between the two "oughts" matters. But more importantly, the question of what to say when there is some degree of doubt about the category a contemplated action falls under is one that has to be addressed by any moral theory. As we noted at the beginning, there is often some degree of doubt about the category an action falls under. This is especially the case when the relevant category is whether or not an action is the intentional killing of the innocent: the probability of innocence very often is a matter of degree. Of course, absolutists are free to insist that their theory is concerned solely with objective-oughts, and so that the discussion (and trouble-making) to come in later sections does not concern them. But if they do this, they turn their backs on an important issue that needs to addressed by anyone of an absolutist stripe. There will need to be a kind of ethical theory that agrees with them in being absolutists about objective-oughts and, in addition, has something pertinent to say about decision-oughts. Call it "absolutism*" if you like, in which case this paper is a discussion of absolutism*, but we will stick with the shorter name. As a matter of record, the "turning the back" stance is not taken by one prominent group of absolutists: namely, those Catholics who oppose early stage abortion while granting that it is far from certain that an early stage foetus is a person. They argue that it is possible that early stage abortion is not the intentional killing of an innocent person because it is not the killing of a person, and, therefore, that it may not be ruled out by absolutist prohibitions against the intentional killing of innocent persons. But, all the same, early stage abortion ought not to be allowed because we cannot be sufficiently confident that the early stage foetus is not a person. Their position is, precisely, that early stage abortion may or may not be something that is objectively wrong but it is certainly something that decision-ought not to be done.

In sum, this paper is concerned with the decision-ought question for absolutist theories, but except where required by clarity, we will drop the prefix "decision" in what follows. Indeed, in everyday life, we typically read the "oughts" we discuss with the prefix taken for granted: as in "If there is doubt about the safety of an inoculation, we ought not make it compulsory" or "It is true in general that we ought not take unjustified risks." These contentions only make sense as ones directed to what decision-ought not to be done. As heralded, our discussion will be largely framed in terms of absolutisms that have an absolute prohibition against intentionally killing the innocent, and we will often refer to the case we now describe.

II. A CONCRETE EXAMPLE

A skier is heading in a direction you know for sure will trigger an avalanche that will kill ten people. You know the only way to save the ten people is for you to shoot him. The probability that the skier intends to trigger the avalanche and kill the ten people is $1-p$. We can agree that our target absolutist theory says it is right for you to shoot if it is certain that the skier intends to kill the ten, that is, if $p = 0$, for in that case you would not be killing someone innocent—you would be protecting the ten in the only way possible from an unprovoked attack. We can agree that our target theory says that it would be wrong for you to shoot if you are certain he simply happens to be skiing in that direction, that is, if $p = 1$, for then you would be intentionally killing someone innocent, and that is never right no matter how many you will be allowing to die by your failure to shoot. The number of lives that would be saved in the example as described is ten, but of course the distinctive position of absolutism is that the number does not matter: it is never right intentionally to kill the innocent no matter how many lives would be saved by doing so. Our question is, What should the theory say for other values of p?

III. THE INFINITE DISVALUE APPROACH

Perhaps the simplest absolutist answer to our question is to hold that whenever there is any chance that an action violates an absolute prohibition, the action ought not to be performed. This is the answer suggested by the absolutists' case against early stage abortion summarized above. In our example, the answer would prohibit shooting the skier whenever there is any chance that he is innocent, whenever, that is, $p < 1$. One way of implementing this answer is to assign infinite disvalue to intentionally killing the innocent and some finite disvalue to allowing people to die. For then the expected disvalue of the shooting—that is, the product of the disvalue of intentionally killing the innocent times the chance that the shooting is an intentional killing of the innocent—will exceed the disvalue of allowing others to die, no matter how many others die and how certain it is that they will die, provided there is some chance that the shooting is indeed an intentional killing of the innocent. It will, on this approach, be impossible to make the action that has some chance of being the intentional killing of someone innocent the right thing to do by making the number allowed to die by refraining from shooting large enough—the numbers allowed to die will be irrelevant, just as absolutists typically say.[6]

The trouble with this response is that there is nearly always some greater than zero chance that someone is innocent. All the evidence may be against them, but induction from the past record of overturned verdicts in cases that looked watertight at the time tells us that there is nearly always some chance that someone who looks clearly to be guilty is in fact innocent. We will get the result that it is never, or hardly ever, right to shoot the skier. Indeed, it will be hard to find any cases where it is right intentionally to kill someone as there is always some chance that the someone is innocent, and a small chance times an infinite disvalue equals an infinite disvalue. We will have a quick (too quick) argument from absolutism against intentionally killing the innocent to an extreme kind of personal pacifism.

The obvious response to this difficulty is to assign a very high but finite disvalue to intentionally killing the innocent. It will then be possible for an intentional killing that is *very* likely the intentional killing of someone guilty to be the right action in the appropriate circumstances. There are two problems with the "very high but finite" disvalue approach, depending on the position taken on the mathematics of disvalue.

---

[6] We get the same result if we work with a utility function that has 0 rather than minus infinity as its lowest value (greatest disvalue) and 1 as its highest value, and assign value 0 to intentionally killing the innocent and some value between 0 and 1 to allowing people to die. For it will of course still be the case that the expected disvalue of the shooting—that is, the product of the disvalue of intentionally killing the innocent times the chance that the shooting is an intentional killing of the innocent—will exceed the disvalue of allowing others to die, no matter how many others die and how certain it is that they will die, provided there is some chance that the shooting is indeed an intentional killing of the innocent. In the one case, the point is that minus infinity times a positive number is minus infinity; in the other that zero times a positive number is zero. A similar point applies to the argument in the next section. We are indebted to the editors.

IV. LARGE BUT FINITE DISVALUE WITH NO BOUND ON DISVALUE IN GENERAL

Any reasonable version of absolutism allows that it is bad to act in a way that leads to death. Refraining from throwing a life belt to a drowning woman is not plausibly excused by saying that it is merely an allowing to die and not an intentional killing. The position is not that there is nothing bad about allowing to die. The absolutist contention is rather that, no matter how bad allowing to die is, it never justifies intentionally killing the innocent.

Now suppose that the way the mathematics of disvalue works is that the disvalue of allowing to die increases as the number that will die increases without any upper limit (that is, the *value* tends to minus infinity). In this case absolutism cannot handle the difficulty created by treating the disvalue of intentionally killing the innocent as infinite by making its disvalue very large but finite. The reason is that by considering a case where enough are allowed to die—we have ten being allowed to die in our skier example as it stands but we could make the number as big as we please—we will be able to construct a case where allowing to die is worse than intentionally killing the innocent. Suppose that the disvalue of intentionally killing the innocent is some large but finite number *N*. Suppose that in our skier example it is all but certain that the skier is innocent so we can set the expected value at almost minus-*N*. All we need to do is make the number certainly allowed to die, *X*, large enough to create a disvalue greater than *N* (that is, a value less than minus-*N*). This result is, of course, unacceptable to absolutists; it is precisely what they deny. It makes the numbers allowed to die critical for what ought to be done.

V. THE BIG BAD NUMBER

Absolutists sometimes in effect refuse to discuss the very hardest of choices: the choices that arise in the horrific cases consequentialists describe when seeking to make the case that there are no defensible absolute prohibitions. According to consequentialists, for every kind of action one might nominate as absolutely forbidden, there are horrific cases where following the prohibition has consequences sufficiently appalling to show that the prohibition should not be given absolute status. One way of thinking about the absolutists' (sometime) refusal to discuss these cases is as embodying a certain view about moral disvalue. On this view there is a big bad negative number *B* which marks a lower bound, a floor, on expected disvalue in the area of action. Things can get worse and worse but can never drop below *B*. On any view like this, absolutists have, it seems, a way out of the objection of the previous section.[7] It need not be the case that making the number of people allowed to die in the skier example big enough can mean that allowing that number to die is worse than that of taking such-and-such a chance that one's action is intentionally killing the innocent. The (expected) disvalue of that chance of one's action being the intentional killing of the innocent may be at the "floor," namely, *B*.

In order to discuss this proposal we need to spell out in a little more detail the way probability and value interact on the proposals we are considering as possible replies by absolutists to our question about how absolutism should handle the problem of uncertainty.

In the skiing example, you have the choice of shooting or not shooting. There are two possible outcomes of shooting: intentionally killing someone innocent and intentionally killing someone guilty (that is, someone who without justification is seeking to kill by triggering an avalanche). There is one possible outcome for not shooting: allowing ten people to die. The expected value of shooting is:

(A) $[p \times V(\text{intentionally killing the innocent}) + (1-p) \times V(\text{intentionally killing the guilty})]$

The expected value of not shooting is:

(B) $[1 \times V(\text{allowing 10 to die})]$.

The problem noted in section III with giving intentionally killing the innocent an infinite disvalue is that whenever p is nonzero, that is, nearly always, (A) will come out worse than (B), which means that it is never right for you to shoot in the skier case—clearly the wrong answer by all lights except extreme personal pacifism, as we noted. The problem noted in section IV with giving intentionally killing the innocent a finite but very large disvalue is that it threatens to make it the case that increasing the number allowed to die in (B) can make (B) worse than (A) even when p, that is, the probability that shooting is indeed the killing of the innocent, is high (in fact, as high as you like). This would mean that we lose the distinctive feature of absolutism, as we noted.

If the big bad number approach is to avoid each of these (disastrous for absolutists) results, it must insist that there is a high probability value, p^*, that has two properties. First, if the probability that the shooting is the shooting of someone innocent is below p^*, then the product of that probability with the value of intentionally killing someone innocent is greater than the big bad number B (the disvalue is smaller). That way we allow the possibility that it is sometimes right to shoot—namely, when what would happen, were you not to shoot, has a value below this value while still being above B. For the big bad number theory must allow that when the probability of innocence falls below a certain value, it is right to shoot. Second, it must also insist that when the "activation" probability, p^*, is reached, the absolute prohibition kicks in. On the big bad number theory, that is to say that when $p = p^*, p \times V$(intentionally killing the innocent) $= B$. The reason is that otherwise we will have the consequentialist-type result that by increasing the number allowed to die from ten to whatever is called for, we can make (B) worse than (A), so making what ought to be done in part a function of how many are allowed to die.

---

[7] Thanks here to Andy Egan but he should not be held responsible for our formulation.

The upshot is that the big bad number approach has to be understood as a *threshold* approach to the question we are posing absolutists. Although it would be wrong to insist that the absolute prohibition on intentionally killing the innocent only comes into play when a contemplated course of action is certain to be the intentional killing of someone innocent—that makes absolutism largely irrelevant in practice as there is almost always some element of doubt—there is, runs the idea, a threshold at which the absolute prohibition kicks in (p^* above); this is the probability level which might be called that at which it is a *moral certainty* that the contemplated action falls under the category in question, as we might put it.[8] In the skiing example, shooting is absolutely prohibited no matter how many would be saved when it is a moral certainty that the shooting is the intentional killing of someone innocent.

We hazard that many absolutists will find this version the most attractive.[9] We now turn to its troubles.

VI. THE THRESHOLD VERSION OF ABSOLUTISM

Any threshold version has a delicate balancing act to perform. Set the bar too high, and the version becomes irrelevant in practice; in real life there is always an element of doubt. Set the bar too low, and you make absolutism about not intentionally killing the innocent too like personal pacifism. But suppose we agree that setting the threshold at 95% is plausible. An obvious question is how an absolutist might give a principled justification for this value. For the difference between 94% and 96% will correspond to a complete shift in ethical outlook on the threshold version of absolutism. At 94% consequentialist considerations can come into play and settle what ought to be done; at 96%, they are absolutely ruled out—no amount of avoided death and suffering justifies intentionally killing someone whose chance of being innocent is 96%. However, we will not pursue this question. But before we leave it, we should emphasize that it would be a mistake to think of the threshold as a function of the available options—or, more precisely, this is not something absolutists can do. A natural thought is that the bigger the numbers allowed to die, the higher we should set the threshold. But that is giving the numbers allowed to die the kind of central role that absolutism precisely rejects.

---

[8] Thanks to Kit Fine here.

[9] It was suggested to Smith by Tony Coady in discussion. We do not know if it is his considered view, however.

The problem we especially wish to press for the threshold view is that it cannot handle actions involving multiple people. We will make the point using a modification of the skiing example. Suppose that there are two skiers, *X* and *Y*, whose probability of innocence is individually just below the threshold, 0.95 as we are supposing, and whose paths threaten separate groups of people. On the threshold view, you ought to shoot *X* if the number saved is large enough because the probability of innocence is below the level at which the absolute prohibition kicks in; ditto for *Y*. But, on the version of absolutism under discussion, you ought not shoot *X* and *Y* no matter how many would be saved, as the probability that shooting both is to shoot at least one innocent person will exceed the threshold, or so we may suppose consistently with their individual probabilities of innocence being below the threshold. This means that the threshold version of absolutism violates agglomeration. It allows the possibility that you ought to shoot *X*, and that you ought to shoot *Y*, when it is false that you ought to shoot *X* and *Y*.[10]

Agglomeration is very appealing and is widely accepted and many will regard this as trouble enough for the threshold version of absolutism. However, we accept the minority position that agglomeration fails in certain special cases. For us the problem is that the threshold view makes the modified version of our skiing example into an implausible kind of moral dilemma. There is no course of action open to you (the potential shooter) which is such that by following it you do all that you ought to do. Here are the courses of action open to you in the example.

---

[10] The difficulty raised here for a threshold version of absolutism is similar to the one raised by the lottery paradox against analyses of a "yes-no" concept of belief, the one sometimes called *acceptance*, in terms of degree of belief reaching a specified high level

(1) Shoot *X* and do not shoot *Y*.
(2) Do not shoot *X* and do shoot *Y*
(3) Do not shoot *X* and do not shoot *Y*
(4) Do shoot *X* and do shoot *Y*.

Here are the verdicts the threshold view gives for each. If you adopt (1), you do the wrong thing by virtue of not shooting *Y* (although you do the right thing in shooting *X*). If you adopt (2), you do the wrong thing by virtue of not shooting *X* (although you do the right thing in shooting *Y*). If you adopt (3), you do the wrong thing by virtue of not shooting *X*, and also the wrong thing by virtue of not shooting *Y*. If you adopt (4), you do the wrong thing by virtue of shooting *X* and *Y*—as there is no way of shooting *X* and shooting *Y* without shooting *X* and *Y*.

This is serious trouble. It is the job of a moral theory to delineate what ought to be done in a way that makes at least one option correct ("at least one" to cover the possibility of ties). We are talking here about the decision-ought of course, and it is part of decision-ought's job description to mark out one or more courses of action as correct (although it may at the same time be the case that every course of action available to the agent is bad). There is of course an interesting debate about whether there are situations in which every available course of action is both correct and incorrect in the decision sense: so-called strong moral dilemmas. (Smith is sympathetic to this possibility; Jackson is not.) However, the situation delivered by the threshold version of absolutism is *not* one where every available course of action is both correct and incorrect. There are two available actions that the threshold version says that you ought to carry out, namely, shooting *X*, and shooting *Y*, but it does not in addition say that you ought not shoot *X* or that you ought not shoot *Y*. Instead it says you that you ought not shoot *X* and *Y*. The problem here is not that the threshold view says that this is a strong moral dilemma; the problem is rather that it puts value and probability together in a way that leads to paradox.

(threshold). What happens if we replace mention of degrees of belief by a "yes-no" notion like acceptance in our discussion of absolutism? The answer is that we change the subject. There is a question for absolutism arising from the often realized possibility that we have one or another nonextreme degree of belief that some action falls under the relevant characterization, and you do not answer that question by discussing another (interesting) one.

There are three things that absolutists might say in reply.[11] First, they may reply that what matters is the status of each action; if each shooting is morally right, no problem if the conjunction is not right. Getting the correct answers action by action suffices. But then it would matter whether you fired two shots or threw a single hand grenade that would kill both *X and Y.* Or maybe the case is one where a single shot could kill both *X* and *Y* by passing in turn through their bodies. We would get the bizarre result that it matters greatly whether you shoot with a single bullet or use two. Second, they may say that there is an available course of action that is right, namely a disjunctive one: shooting *X*-or-*Y*-but-not-both. The trouble for this suggestion is that there exactly two ways of shooting *X*-or-*Y*-but-not-both—by shooting *X* and not shooting *Y,* and shooting *Y* and not shooting *X,* and both involve doing something wrong (not shooting *Y* in the first case, and not shooting *X* in the second). Third, they may say this is one of the cases where what ought to be done is a function of what else is done or not done. There are cases where you ought to do *A* and *B,* but if you do not do *A,* then you ought not to do *B.* A dentist ought to administer a local anaesthetic and then start drilling, but if they have not administered the anaesthetic, then they ought not to start drilling. But the trouble we have made for the threshold version of absolutism is that it has the consequence that in our two skiers' case, you ought to shoot *X* *independently* of whether or not you shoot *Y,* and conversely; at the same time, you ought not to shoot *X* and *Y.*

VII. ABSOLUTISM AS A THESIS ABOUT THE MIND IN ACTION

We have been treating absolutism as absolutely forbidding certain *actions,* but sometimes it is thought of as a thesis about the mind in action, a thesis about which *intentions* a moral agent can properly form and act on. In terms of our target version of absolutism, the contention is not that the *action* of intentionally killing the innocent is absolutely forbidden but that *forming the intention* of killing the innocent is.[12] When it is said that it is *never* right intentionally to kill the innocent, what is meant is that it is never right to form that intention. The suggestion is not, of course, that there is nothing to say about action *per se*;

---

[11] Thanks here and in the previous paragraph to the editors.

[12] As in some versions of the principle of double effect, see, for example, Anscombe, "War and Murder," in *Ethics, Religion and Politics,* chapter 6. The most worked-out version of a "mind in action" ethical theory we know of is Michael Slote's "agent-based" approach in *Morals from Motives* (New York: Oxford, 2001); however, he is not defending absolutism. Slote has detailed discussions of who can and who cannot be read as agent-based theorists. Thanks to Dave Sobel for discussion here.

indeed, the precise relation between the moral status of the intention and of action in accord with it is open to debate in such theories.[13]

There are troubles on two fronts for this suggestion. One arises from the possibility of uncertainty about one's intentions, the other from asking after the content of the intention.

Although it has been suggested to us in discussion that absolutisms of the kind in question avoid the troubles we have been raising on the ground that subjects can be certain of what they intend, this is a mistake. We do not have incorrigible knowledge of our mental states and the point is especially obvious and widely agreed in the case of intentions (and the same goes for motives). It is a commonplace that we can believe that our intention is to serve someone's interests by informing them of an unpalatable truth they need to know, when it is obvious to our friends that our real intention is to upset them. Likewise, subjects can sometimes be wrong about an action's being in self-defense, or about their "real" intention in making a devastating criticism in discussion. The problems we have been raising for absolutism about actions arising from uncertainty about their nature can be raised *mutatis mutandis* for absolutism about intentions.

The second problem arises from asking after the content of the intention. How might the absolutism in question—absolutism thought of as forbidding forming the intention to kill the innocent—specify the forbidden intention?

One possibility is that the forbidden intention is specified as forming the intention to kill someone who is innocent, without any reference to whether or not the agent has any opinion about whether the "someone" is or is not innocent; it is no part of the specification of the forbidden mental state that the agent has a view about the innocence or otherwise of the person or persons about whom the intention to kill may be formed.

This is a very unattractive option. Suppose that *A* is in fact innocent. It would be just as forbidden to form the intention of killing *A* when you have every reason to believe *A* guilty or every reason to believe *A* is attempting to kill you, as it would be to form the intention in a case where you had every reason to hold *A* innocent. Moreover, this specification would mean that absolutism as a thesis about intentions would conspicuously fail to address the decision-ought question we discussed earlier. An agent wondering whether or not it is permissible to form the intention to kill *A* is inevitably faced with that question

---

[13] See, for example, the discussion of Kantian theories in C.D. Broad, *Five Types of Ethical Theory* (London: Routledge Kegan Paul, 1930).

in the context of one or another view about A's innocence or guilt. Perhaps the agent's view is that A is certainly innocent, or probably guilty, or probably about to attack, or...?

The upshot is that absolutism's specification of the forbidden intention must include something about the agent's beliefs about the putatively intended victim. When you ask yourself what intention you ought to form according to absolutism in our skier example, your opinion about the skier's innocence has to enter the picture. What must be absolutely forbidden is forming intentions to kill people about whom the agent has such-and-such beliefs concerning their innocence. But now we will see is that the various possible spellings out of this schema raise problems broadly analogous to those raised in previous sections. Suppose, first, that the rule is:

> R': it is absolutely forbidden to form the intention to kill someone you believe with degree of belief one to be innocent.

It is hardly ever the case that you will have such a degree of belief. Indeed, on certain views about degrees of belief, a degree of belief of one can only be possessed, or only possessed justifiably, when the belief is *a priori* true. Certainly, it will never be the case that you have such a degree of belief in the skier or like examples; there will always be some chance that he is guilty. But then R' fails to rule out any real-life intentions.

Suppose, second, that the rule is:

> R'': it is absolutely forbidden to form the intention to kill someone you believe with degree of belief p^* or above to be innocent.

R'' is a "mind in action" version of the threshold view we discussed earlier and is exposed to the same problems. What principle sets the right value of p^*? And in cases of actions involving more than one person, we will get cases where the subjective probability that one or another person is innocent is above the threshold but the probability of each person being innocent is below the threshold.

Suppose, third, that the rule is:

> R''': it is absolutely forbidden to form the intention to kill someone you believe with any degree of belief above zero to be innocent.

In this case, it will never be right to shoot the skier and, more generally, as there is always in real-life cases some chance that the person you intend to kill is innocent, R''' will imply that it is never right to form the intention to kill. Absolutism will be reduced in practice to a kind of personal pacifism.

A familiar theme in the philosophy of action is that an action can be intentional under one description but not under another. Killing the man behind the arras may be the same action as killing Polonius, but it is only intentional under the description "killing the man behind the arras."[14]

Is there a halfway house for absolutism that sees it as a thesis not about intentions *per se*, and not about actions *per se*, but about actions under intentional descriptions, that avoids the troubles we have been raising?[15] As far as we know, there is no extant version of absolutism framed in this style but it might run somewhat as follows, framed in the terms of our target example:

> R^{iv}: It is absolutely forbidden to perform an action under the description "intentionally killing someone innocent."

But what, exactly, is R^{iv} forbidding? On the face of it, actions under certain descriptions, but what kind of animal is an action under a description?

The obvious answer is that an action under a description is simply an action that falls under the description. What is absolutely forbidden are certain actions that have the property of falling under the description "intentionally killing someone innocent." But does it matter whether or not the agent has any opinion about whether the "someone" is or is not innocent? To say it does not matter is to say that an action of killing someone you have every reason to believe guilty is as forbidden as killing someone you have every reason to believe innocent in the case where they are in fact innocent. Some will bite this bullet but the really important problem is that the halfway position so understood fails to address the decision-ought question we are concerned with. An agent wondering what to do is inevitably faced with that question in the context of one view or another about innocence or guilt. If absolutism is to address the decision-ought question, it must say something about actions forbidden under descriptions of the form "intentionally killing someone believed innocent to such-and-such a degree." (How does one act on a stock tip like "Put your money in insurance stocks"? By putting one's money in stocks *believed* to one degree or another to be insurance stocks.) But once degree of belief enters the picture, as it must, the same issues *mutatis mutandis* will arise as arose for R', R'', and R'''.

How else might we understand R^{iv}? The view that one and the same action can be intentional under one description and fail to be intentional under another is the view that being intentional is not a property of actions; instead being intentional under D is, for various D. Otherwise we would have a violation of Leibniz's law. Perhaps R^{iv} is in effect the view that actions do not have the property of being absolutely forbidden: not in the sense that strictly speaking intentions have that property—we have seen the serious problems with that position already—but in the sense that actions have the property of being absolutely-forbidden-under-the-description-"intentional killing of someone innocent." But now we have a theory that says nothing about which actions, and nothing about which intentions, are absolutely forbidden. Of course we might restore the needed property of delivering answers as to what is absolutely forbidden by allowing *detachment*. Detachment would say that from "It is absolutely forbidden to perform an action under the description 'intentionally killing someone innocent'," and "This action falls under the description 'is an intentional killing of someone innocent'," it follows that the action is absolutely forbidden. But now we are back with the problems rehearsed above. We need to ask questions like, "How likely is it that the action falls under the description?" and, "What degree of belief does the agent have that the person is innocent?" We are back where we started.

---

[14] Donald Davidson, "Agency," in *Essays on Actions and Events* (New York: Oxford, 1980), pp. 43–61.

[15] Thanks to Bob Goodin for convincing us of the need to address this position. It is not one he holds.

VIII. ARE WE PRESUMING CONSEQUENTIALISM'S CONCEPTUAL FRAMEWORK?

A natural thought is that we are setting our discussion in a framework that biases against absolutism. Are we not approaching examples like the skiing one in terms of a choice between the *option* of shooting and not shooting, and asking about the probability that the shooting would have as an *outcome* the killing of someone innocent? And is the conceptual apparatus of options and outcomes not the apparatus characteristic of consequentialism? But in fact the puzzle we have raised does not depend on thinking of the shooting as distinct from intentionally killing someone innocent in the case where the shooting is in fact of someone innocent. We do not have to think of the killing of someone innocent as in any sense the outcome of the shooting; "they" can be one and the same action. Take the way we put the problem in schematic form earlier:

> Our concern will be with the following kind of question: agent S has to choose between doing A, where A has p chance of being K and $1-p$ chance of being not-K, and doing B, where B is certain to have such-and-such consequences. What should a theory that absolutely prohibits doing a K action say about what happens to what S ought to do as p rises and falls?

Our problem is generated not by a bifurcation between action and outcome—we can be completely neutral on that subject—it arises from a point about the epistemology of action rather than its metaphysics. We can be and often are in doubt about whether some action does or does not fall under the description that the absolutist regards as crucial for its moral status. That is what generates our problem, not a potentially controversial metaphysics of action.

FRANK JACKSON

The Australian National University

MICHAEL SMITH

Princeton University

WHAT IS WRONG WITH WEAKNESS OF WILL?*

Why has weakness of will attracted so much philosophical attention? Well, it might be said, anyone who is not a skeptic about the possibility of weakness of will can see that we must leave room for it when describing the rational processes underlying practical deliberation. If the transition from thinking it best to do *x* to forming the intention to do *x* is thought to be automatic, or if the step from intending to do *x* to doing *x* when the moment arrives is taken to be inevitable, it will seem impossible for practical reasoning to stall in the manner characteristic of weakness of will. This is certainly true, but similar remarks could be made about the need to leave room for other defects in the connections between deliberation, motivation, and action. We procrastinate by failing to determine that the time to act has arrived. We form unreasonably ambitious future intentions in order to assuage a current sense of guilt. We impulsively change our minds about what we ought to do in order to rationalize doing what we are most inclined to do. These interesting aberrations have gone unremarked while weakness of will gets all the attention. Why?

Many would say that unlike those other failures, which can be difficult to recognize, weakness of will wears its rational defect on its sleeve. Whenever we judge that it would be best not to do *x*, while intentionally doing *x* without relinquishing this judgment, we condemn quite explicitly the intention on which we act. This observation gives rise to the attractive idea that *weak-willed agents indict themselves of irrationality* as they fail to comply with their own practical judgments.[1] Weakness of will is often treated as the paradigm of practical irrationality just because of the explicit and quasi-logical form of incoherence in the weak-willed or *akratic* agent's attitudes.

---

*For comments on earlier versions of this essay I am grateful to audiences at philosophy department colloquia at the University of Minnesota at Morris, the University of California at Los Angeles, and Clark University, and to audiences and fellow participants at the conference on Practical Irrationality at the University of Montreal, and the conference on The Will in Moral Psychology at the University of Edinburgh.

[1] T.M. Scanlon distinguishes between failing to do what one has most reason to do and failing to do what one believes that one has most reason to do and argues that only the latter should count as irrationality in *What We Owe to Each Other* (Cambridge: Harvard, 1998). Agents may be open to rational criticism for failing to see that they have reason to do something, but irrationality must involve a defect more flagrant than this. What Scanlon calls "irrationality in the clearest sense" is the kind of irrationality that

0022-362X/06/0306/284–311

Underpinning this idea is the assumption that practical rationality can be characterized as a kind of executive capacity that produces congruence between judgment and action by ensuring that practical judgments are faithfully enacted.[2] Some are happy to call this capacity the will, so that the label "weakness of will" is literally true.[3] Others who speak more broadly about motivational capacities nevertheless characterize practical rationality or practical reason as the correspondence of motivation and action with judgment.[4] I will call this conception of the scope and function of practical rationality the *Obedience View.*

The Obedience View assumes that practical rationality consists of the absence of the kind of irrationality displayed by weak-willed agents. Yet this assumption severely limits the resources available for characterizing what is wrong with weakness of will. Consider how fruitful it would be to pursue a comparable strategy for characterizing politeness. First, one would identify deliberate rudeness as the paradigm of impoliteness, pointing out that since deliberately rude agents are rude by their very own lights, they indict themselves of impoliteness. Then one would proceed to define politeness as the absence of that defect. Such an approach could not account for the fact that clumsy and inattentive agents fail to be polite by being unwittingly impolite. Doubtless there is something to be learned by thinking about agents who fail to meet standards that they themselves acknowledge, but such reflection should not rule out the possibility that there might be other less self-conscious ways of failing to meet those standards.

In what follows I will develop an account of the practical defects in weakness of will that implies that the immediate and internal form of

---

occurs "when a person's attitudes fail to conform to his or her own judgments" (p. 25). I will follow this usage and reserve the term *irrational* for especially egregious rational failures.

[2] Philip Pettit and Michael Smith emphasize the importance of identifying a distinctively practical form of rational failure in "Practical Unreason," *Mind*, CII (1993): 53–79.

[3] R. Jay Wallace, "Three Conceptions of Rational Agency," *Ethical Theory and Moral Practice*, II (1999): 217–42. Robert Audi observes that one reason for taking incontinent actions to be irrational is that "they seem to manifest a *malfunction*: the will is not carrying out its proper function—to produce conformity between action and practical reason"—"Weakness of Will and Rational Action," chapter 12 of *Action, Intention, and Reason* (Ithaca: Cornell, 1993), p. 324.

[4] In "Practical Unreason," Pettit and Smith describe rational action as action in which motivation conforms to the results of deliberation not as a result of blind obedience, but through sensitivity to the values that lead to choice: "The values that weigh with the agent in deliberation serve also to arouse a desire for the option which they deliberatively support. Their net impact on arousing desire—their net desiderative force—corresponds to their net deliberative weight" (p. 57).

incoherence in the weak-willed agent's attitudes is not what makes weakness of will a failure of practical rationality. Instead, weakness of will involves two central defects, one procedural and one substantive, each of which can occur even when agents are not acting against their better judgment. First, weakness of will is a failure to be resolute: a failure to carry out a special kind of intention which can be called a *resolution*.[5] Rationalizing one's reluctance to act by revising one's intention is also a failure to be resolute, and the defect in rationalizing and the defect in weakness of will can be given a common explanation. Second, weakness of will is a failure to do what one has most reason to do. Typically agents who act against their better judgment fail to do what they have most reason to do, but this is not always the case. In the section that follows, I will lay the groundwork for this expanded account of practical rationality by arguing that an adequate account of what is wrong with weakness of will must allow that a lucidly akratic agent—one who does not abandon the judgment that she fails to carry out—may be doing better from the standpoint of practical rationality than an agent who rationalizes her unwillingness to act by abandoning her original practical judgment.

I. WEAKNESS OF WILL AND PROBLEMS FOR THE OBEDIENCE VIEW

Let us juxtapose an agent who revises a practical judgment in order to rationalize a failure to be motivated by it —a *rationalizing agent*— and an akratic agent who fails to comply with a judgment while continuing to hold it. Reading a novel late at night you decide that it would be best to stop at the end of the chapter since it is more important to be rested the next day than to keep on reading. But the novel is gripping and when the end of the chapter approaches, you read on. You have rationalized if you caved in and decided that it would be best to read on as you did so. You were akratic and weak-willed if you continued to believe that you should stop at the end of the chapter and failed to get yourself to do so.

Since agents who rationalize a change of mind about what would be best might well preserve perfect congruence between judgment and action, the Obedience View implies that the defect displayed by agents who rationalize is entirely distinct from the defect displayed by the weak-willed. Yet what is wrong with the rationalizing agent seems closely akin to what is wrong with the weak-willed agent. First of all, they can be very hard to tell apart, even when the agent is oneself. (Did you retain the judgment that it would have been best to stop at the end of the chapter as you read the first sentence of the next chapter? Or had you already revised your judgment about what would be best by then?) Second, the rationalizing agent is deficient in some central respect involving motivation. If the defect in weakness of will involves a failure to contain recalcitrant motives, then rationalizing agents seem to display it also. Yet rationalizing need not involve weakness of will: the rationalizing agent might revise her judgment about what is best sufficiently far in advance so that she does not, even for a moment, flout the earlier judgment in action. If the Obedience View is correct and practical rationality is simply compliance with evaluative judgment, the rationalizing agent shows no lapse in practical rationality since compliance is perfectly preserved. (By changing her mind, the rationalizing agent has neatly disposed of the evidence she would require to indict herself of irrationality.) This verdict seems to conflict with the intuitive thought that if the rationalizing agent had displayed more strength of will, she would not have succumbed to the temptation to revise her judgment about what is best.

---

[5] Similar accounts have been provided by Thomas Hill, "Weakness of Will and Character," in *Autonomy and Self-respect* (New York: Cambridge, 1991), pp. 118–37; and Richard Holton, "Intention and Weakness of Will," this JOURNAL, XCVI, 5 (May 1999): 241–62.

Those who hold the Obedience View might say that this is no objection to their view since rationalizing involves a defect in a different department, theoretical rationality, and need not involve a lapse in practical rationality. An agent who rationalizes a failure to do what she had intended by changing her mind about what would be best does not have good reason to do so, since she revises the judgment on the basis of something—her failure to be motivated to comply with it—that has no evidential value. By contrast, they might remark, a lucid weak-willed agent who does not lose her grip on what would be best because of her reluctance to act, does not display this defect in theoretical rationality, but because she fails to be motivated to act on the judgment she currently holds, she displays a form of practical irrationality. The two failures might be causally linked—an akratic practical failure might prompt the rationalizing theoretical failure—but the defects displayed are utterly distinct.

Yet when we compare the rationalizing late-night reader and the akratic late-night reader, the akratic seems to be doing better than the rationalizer in containing her reluctance to stop reading. In some respect concerning motivation, she is doing better, yet the obedience view says that she is doing worse. Furthermore, the lucidly akratic agent's resistance to the temptation to rationalize a change of mind seems to be just the kind of thing that generally produces strength of will. If resisting the urge to rationalize is just displaying strength in the domain of theoretical rationality, and if people are often strong-willed because they resist the urge to rationalize, then strength of will must be counted as a form of theoretical rationality while weakness of will must be counted as a form of practical irrationality. This seems absurd.

Furthermore, according to the attenuated conception of practical rationality that is implied by the Obedience View, if you agree with Fred that he ought to get more sleep, you and Fred agree in your theoretical judgments. If you join Fred in considering which means he should adopt (sleeping later? retiring earlier? taking naps?) this too is a question in the province of theoretical rather than practical rationality. Yet we would ordinarily call *practical* in a broad sense any form of deliberation about which means Fred should adopt. And we would ordinarily call *practical* in a more central sense reflection about which means would be most likely to be complied with, on the assumption that Fred will find his sleeping habits hard to change, and reflection about which method would yield the most additional sleep with the least effort. How likely it is that Fred will succeed seems to be a matter of how feasible—how practical—his newly drafted plans are. If practical rationality is concerned with Fred's formulation of his plan to get more sleep and his success in executing it, then how well drafted his plan is and how feasible it is fall within the purview of practical rationality. These are the methods by which Fred engineers his compliance. If all goes well, he will get more sleep without having to undergo dramatic tests of willpower in which he is strongly tempted not to carry out his resolutions. But whether things go well has as much to do with the reasonableness and effectiveness of his plans as it has to do with his disposition to be motivated to do what he believes he ought to do, and the former have as much claim, and perhaps greater claim, to be a form of reason than the latter.

Proponents of the Obedience View often assume that practical rationality is a form of practical reason and that it involves a form of motivational compliance that is rationally guided. Yet it is hard to find anything that answers to that description in our psychological constitutions. This difficulty is confronted by John Broome who asks, "How should we characterize the right way for a normative belief to explain an act?" He asserts "If reason is to be truly practical, the right way should somehow involve reason," but he is troubled that a natural disposition to do what one believes one ought to do does not involve reason.[6] I suggest that the solution is to abandon the assumption that the step from judgment to action must be anything like a form of inference in order to constitute practical rationality. If one takes a broader view and allows that practical rationality might be constituted by a variety of capacities and dispositions—some quite reflective and others quite unreflective—that support practical agency, then the main rationale for restricting practical rationality to motivational compliance with judgment will disappear.

---

[6] Broome, "Reason and Motivation," *Proceedings of the Aristotelian Society*, Supplementary Volume LXXI (1997): 131–46.

The view that I have sketched must confront an immediate and obvious challenge. If weakness of will and rationalizing are both defective simply because they are failures to be resolute, then what distinguishes a failure to be resolute from reasonably changing one's mind about what is best and revising one's intention accordingly? After all, revising an intention is not in itself irrational, and rationalizing can be quite hard to distinguish from a reasonable change of mind. In fact, rationalizing typically masquerades as a reasonable change of mind.

One possible line of response to this challenge is to emphasize the fallibility of an agent's judgment and to defend the claim that there is nothing procedurally wrong with cases of weakness of will or, in general, failures to be resolute. Instead, it might be maintained, in each case, the rationality or irrationality of carrying out an intention depends on its substantive credentials—on whether the agent has most reason to carry it out.[7] To take this route is to give up on the idea that there is typically some kind of *procedural* defect in weakness of will that constitutes a failure of practical rationality, and to rely instead on the verdict that if there is a defect in weakness of will, it can be accounted for on purely substantive grounds: the weak-willed agent fails to do what she has most reason to do. I will argue that practical rationality does have structural or procedural elements, just as theoretical rationality does, and should not simply be identified with a complete account of what we have good reason to do.[8]

---

[7] See Nomy Arpaly, "On Acting Rationally against One's Better Judgment," *Ethics*, cx (April 2000): 488–513; Audi, "Weakness of Will and Rational Action"; and my paper "Is Weakness of Will Always Irrational?" in Owen Flanagan and Amelie Rorty, eds., *Identity, Character, and Morality* (Cambridge: MIT, 1990), pp. 379–400. Hill provides a refreshingly broad characterization of weakness of will in "Weakness of Will and Character" which would include rationalization as a manifestation of it, but his account of the defect in weakness of will makes it depend upon the irrationality of making or breaking the resolutions involved. A somewhat similar distinction between procedural and substantive dimensions of rationality is drawn by Derek Parfit in "Reasons and Motivation," *Proceedings of the Aristotelian Society*, Supplementary Volume LXXI (1997): 98–130.

[8] This contrasts sharply with Allan Gibbard's use of the term "rational" in *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge: Harvard, 1990) as a term of art to express the sense in which "what it is rational to do settles what to do" (p. 49). Gibbard

If we say that it is rational to be strong-willed only when the judgments lying behind our intentions accord with what we have good reason to do, then we would have to know or guess at the true reasonableness of these judgments in order to know whether we should act on them. We would be placed in the condition of athletes told to follow the advice of their coaches only when their coaches have offered good advice—a condition in which a tendency to follow the coach's advice would be appropriate only when it would be entirely unneeded. Nevertheless, we should not assume that these structural and procedural elements in practical rationality are such that violating them necessarily counts as irrationality. If practical rationality involves a variety of goals that might, in some circumstances, be in tension with each other, then failing to satisfy one of these goals might fall short of constituting outright irrationality.

Familiar examples of weakness of will display three notable defects: (1) the failure to do what one has judged it best to do, (2) the failure to do what one has most reason to do, and (3) the incoherent attitudes of agents who fail to do what they believe they ought to do. In the next section, I will distinguish between the first defect, which I call a failure to be resolute, and the third defect, the presence of akratic attitudes, and show that each can occur without the other. In sections III and IV, I will argue that the failure to be resolute is the central procedural defect in weakness of will, and that failures to be resolute need not display the second feature, the failure to do what one has most substantive reason to do, or the third feature, the presence of rationally incoherent akratic attitudes. Sections V and VI distinguish weakness of will from other more benign failures to be resolute by gauging the kind of substantive failure involved and identifying the conditions in which resoluteness in carrying out an intention can be presumed to be more rational than revising it. Section VII will draw on these elements to provide a more detailed argument for my initial claim that an akratic agent who holds on to a judgment that is not being carried out may be doing better, from the point of view of practical rationality, than a non-akratic agent who revises the judgment or simply abandons it in the face of motivational failure.

II. WEAKNESS OF WILL AS A FAILURE TO BE RESOLUTE

Richard Holton has contested the standard view that weakness of will consists in action against one's better judgment. Reserving the term

---

acknowledges that other phrases, like "the best thing to do" or "the thing to do" might capture this notion better when it is applied to questions of action (p. 49).

akrasia for action voluntarily undertaken against one's best judgment, Holton argues that weakness of will need not involve akrasia (*op. cit.*, p. 243). He observes that a smoker who has resolved to quit can start smoking again "blithely" without believing, at that time, that not smoking is best (*op. cit.*, p. 254). This failure to be resolute is a form of weakness of will he claims, but since the agent has abandoned his judgment that quitting is best, it is not akrasia.[9] He suggests that "the central cases of weakness of will are best characterized not as cases in which people act against their better judgment, but as cases in which they fail to act on their intentions.... Weakness of will arises" he says "when agents are too ready to reconsider their intentions" (*op. cit.*, p. 241).

Holton is absolutely correct, I believe, to claim that the central defect in weakness of will can be characterized in a way that does not involve akrasia, or action against one's current judgment, as a necessary element. We might sometimes retain the evaluative judgment that we are flouting when we succumb to temptation, and we might sometimes hear the voice of self-reproach as we fail to carry out some resolve, but this is not what constitutes the central defect in weakness of will.[10] The smoker trying to quit who lights up and the dieter sneaking dessert need not act against what they currently hold to be best; the defect they manifest consists in a failure to be resolute in carrying out a prior intention.

The fact that akratic failures to be resolute and non-akratic "rationalizing" failures to be resolute can be hard to distinguish, even in ourselves, raises a problem for the view that weak-willed agents indict themselves of irrationality. If agents themselves are not sure how to describe any particular episode, then they know only that they indict themselves of irrationality if they still hold the judgment that they flout, but not otherwise. This sounds more like an incentive to rationalize than a condemnation of weakness of will. A view that can condemn either possibility, rationalization or akratic weakness of will, without needing to distinguish them clearly has a significant explanatory advantage.

---

[9] In "Weakness of Will and Character," Hill describes weakness of will in even broader terms. On his account it is a character trait that manifests itself in a failure to make resolutions or to persevere in carrying them out: "weak-willed persons are the sort of persons who repeatedly and more than normal (a) make inadequate efforts, (b) resolve with too little determination, (c) break their resolutions, with or without a struggle, and/or (d) too readily abandon their undertakings" (p. 130).

[10] Simon Blackburn asks: "Isn't it more plausible to say that when we are weak-willed we change our mind, perhaps for bad reasons, about what we ought to do? In fact, the better the grounds we have for calling an episode of weakness of will voluntary and non-compulsive, the more it looks like a change of mind"—*Ruling Passions: A Theory of Practical Reasoning* (New York: Oxford, 1998), p. 191.

Holton argues for a second important claim: that agents can display akrasia without displaying weakness of will. He considers a situation from Kingsley Amis's novel *Lucky Jim*: Dixon believes he should tell his hostess that he made cigarette burns in the bedclothes and a nearby rug during the night (he was smoking in bed while drunk), but his hostess is the department head's wife, and because he believes that he will not be able to bring himself to confess what he has done to her, he does not form the intention to do so. He is not going to do what he thinks it would be best to do, so he is akratic, but he does not fail to be resolute, Holton argues, because he never forms the intention to tell her (*op. cit.*, p. 258). A failure to be resolute must involve a failure to carry out an intention, Holton specifies, so there can be akrasia without a failure to be resolute.

Taking up Holton's suggestion would allow us to distinguish between three possibilities. (1) Agents might display akrasia by failing to form an intention to do what they believe that they ought to do (for example, by believing that they ought to get more exercise without forming an intention to get more exercise). This is akrasia without a failure to be resolute. (2) Agents might display akrasia as they fail to be resolute in carrying out their intentions (for example, by failing to carry out an intention to exercise today while still believing that one should exercise today). (3) Agents might fail to carry out their intentions without displaying akrasia because they have revised the original judgment that led them to form the intention (for example, by failing to carry out an intention to exercise today after deciding to make an exception for today or after simply abandoning the plan without attempting to justify doing so). In the next section, I will discuss Holton's account of the defect in cases (2) and (3), the akratic and the non-akratic failures to be resolute; in section IV, I defend an alternative account of the procedural defect they display.

III. A PRAGMATIC ACCOUNT OF THE IRRATIONALITY OF WEAKNESS OF WILL

Holton argues that "actors show weakness of will when they revise an intention as a result of a reconsideration that they should not have performed; that is, when their reconsideration exhibits tendencies that it is not reasonable for the agent to have" (*op. cit.*, p. 248). Appealing to Michael Bratman's view,[11] which cites pragmatic grounds for agents to be disposed to favor the stability of intentions once they are formed and to disfavor the reconsideration of intentions, Holton locates the irrationality of weakness of will in the agent's possession and exercise of dispositions that it is not reasonable to possess. He observes that even if it would have been in one's interest to revise one's intentions on some particular occasion, given the merits of the particular case, the tendency that led one to do so might nevertheless be a tendency that, over the long term and regarding one's handling of intentions in general, it would not be rational to have. His account implies that

---

[11] Bratman, *Intention, Plans, and Practical Reason* (Cambridge: Harvard, 1987).

> sometimes it will be rational not to reconsider an intention even when reconsideration would, in that instance, have been in the agent's interests. In general, once we have decided in which restaurant to eat, it is a good idea to let the matter rest, without endlessly discussing the pros and cons; and this is true, even though it occasionally means that we shall go to a restaurant that we could have realized was not the best choice (*op. cit.*, p. 248).

On Holton's view, a revised judgment that yields a substantive improvement might nevertheless be a procedural failure. It might be quite difficult to judge whether a particular failure to be resolute displays this kind of defect since we would have to know whether it is reasonable for us to possess the dispositions and tendencies that produced the episode. Individuating those tendencies and assessing them on pragmatic grounds would constitute, I suspect, a more complicated task than Holton acknowledges.

It might involve assessing how specific tendencies interact with a variety of other tendencies, since there is considerable variation in how people make decisions, without clear differences between them in rationality. As individuals we vary quite a lot in such dimensions as our susceptibility to regret, our risk aversion, our willingness to revise plans once they have been made, and our ability to improvise when our plans are incomplete or imperfect. And even for individuals who are quite similar in these deliberative traits, there may be more than one way to strike an equilibrium between them so as to have a coherent deliberative style that makes efficient use of deliberative resources like time, attention, and memory. How we strike a balance between the need to plan and the need to think about other things can influence how much reason we typically have to be resolute in carrying out our intentions. People who have decided swiftly but are willing to elaborate their plans as time passes and are quick to revise their intentions when the need arises may be doing neither better nor worse than people who have deliberated carefully and then are reluctant to revise their intentions. There may be no uniquely advantageous threshold for intention revision.

Even if we grant the simplifying assumption that there are some traits that are common to any acceptable way to balance such concerns, and other traits so disastrous that they would make equilibrium or stability impossible, the perspective that is required to identify those traits would be that of a cognitive engineer studying the various mechanisms that together set an appropriate threshold for intention revision in human agents. But that is not our perspective on ourselves. If the standard for procedural practical rationality is set by considering which dispositions it would be ideally rational or even minimally rational for us to possess, then we might remain forever in the dark about which, if any, of our akratic failures to be resolute counted as weakness of will.

In later work, Holton has argued that it is by ensuring that reconsideration does not occur and that reflective processes are not fully brought to bear that agents can best maintain their resolutions.[12] Holton assumes that rationalization will seem like rational deliberation once it is underway, so it is more effective to use nonreflective means to put a block on reconsideration to keep this form of deliberation from being triggered. It may be true that successfully resolute people do not reconsider their intentions, but it does not follow that they succeed in being strong-willed because they are governed by strong nonreflective habits that block nonreconsideration. Resistance to the urge to rationalize may be more effective if it is reflective in nature, because it must be capable of discounting and doubting the kinds of considerations that insistently present themselves as good reasons to consider changing one's mind, though they are not in fact good reasons.[13] In fact, those situations in which we stoutly resist the temptation to reconsider our plans are situations in which we already have reason to suspect that compliance might be hard to engineer. This could be a stance adopted for a reason, not a general nonreflective disposition it is rational to possess.

It is important to emphasize, as Bratman and Holton do, that we have finite deliberative resources and that we must construct models of rationality that do not presuppose that deliberation is costless and that it is always rational to optimize our plans. But at the same time, we should not lose sight of the fact that there is a kind of economy in the way in which many of our intentions are structured and assessed with continued sensitivity to the kinds of reasons we have for adopting them. In remaining aware of our reasons for forming certain intentions, we simplify the task of determining what counts as a reason to reconsider or revise them. If your intention to take a walk grew out of a desire for fresh air, then changing your destination on a whim just adds a pleasant feeling of spontaneity; if you were walking to the train station to catch the last train home, a sudden urge to change your destination is a sign of something more subversive.

---

[12] Holton, "How Is Strength of Will Possible?" in Sarah Stroud and Christine Tappolet, eds., *Weakness of Will and Practical Irrationality* (New York: Oxford, 2003), pp. 39–67.

[13] Scanlon (*What We Owe to Each Other*, chapter 1) characterizes the influence of desires on deliberation by describing them as tendencies to take oneself to have reason to act in a certain way or tendencies to view certain considerations as reasons to act. I doubt that the motivational role of desires is constituted by such tendencies, but it is certainly true that desires typically give rise to them.

IV. WHY WEAKNESS OF WILL NEVER TAKES US COMPLETELY BY SURPRISE

Anticipating the objection that his account does not distinguish weakness of will from other forms of unreasonable irresoluteness like vacillation and caprice, Holton supplies an amendment: "We can think in terms of a genus with two species. The genus is irresoluteness: unreasonable revision of intention. The two species, weakness of will and caprice, are distinguished by whether the intentions revised are contrary inclination defeating intentions or not."[14] The weak-willed agent, unlike the merely capricious one, revises an intention that was formed at least in part in order to defeat anticipated temptation. Holton explains: "one reason for forming an intention can be in an attempt to overcome contrary desires that one believes one will have when the time comes to act. An intention can be, as we might say, *contrary inclination defeating*" (*ibid.*, p. 250).

This idea, severed from Holton's project of providing pragmatic grounds for calling weakness of will irrational, yields all by itself a criterion for distinguishing failures to be resolute from more benign changes of mind. We can define a special sort of intention, *a resolution*, as an intention to act despite anticipated contrary inclination that is formed because the agent anticipates that she will feel the pull of contrary inclinations when the time to act comes.[15] A failure to be resolute is a case in which one of these contrary inclination defeating intentions is defeated by contrary inclination of the sort anticipated. Thus, a failure to be resolute can be characterized as a certain kind of sequence: first, the formation of the judgment that one ought to do *x* despite anticipated contrary inclination; second, the formation of an intention to do *x* despite the anticipated contrary inclination (the resolution), and third, a failure to carry out the intention because of contrary inclination of the sort anticipated.[16]

---

[14] "Intention and Weakness of Will," p. 251.

[15] In "How Is Strength of Will Possible?" Holton calls the contrary inclination defeating intentions *resolutions*. Hill also counts a failure to carry out a resolution as a form of weakness of will, even if the agent has revised her judgment, observing that "resolutions (in contrast to mere plans and policies) would be pointless if they did not direct one not to reconsider or deviate from the resolve for certain anticipated reasons (for example, the arrival of an expected temptation or pressure)"—"Weakness of Will and Character," pp. 131–32.

This allows a failure to be resolute to be constituted by a change of mind, but also distinguishes this form of change of mind from other, more benign ways of changing one's mind. If we are confident in ascribing these attitudes to an agent, then we can ascribe a practical defect. One need not be able to adopt the perspective of a cognitive engineer to see that a technique of self-management, one that partly constitutes practical rationality for adult human agents, has failed.[17]

Mature human agents form such resolutions because they are experienced enough to realize that weakness of will does not come about simply because reason cannot compete with desire. A failure to be resolute is often the result of influence by considerations that have led one to revise an earlier intention (as in akrasia) and perhaps also the judgment that gave rise to it (as in rationalization). This occurs when agents succumb to the temptation to give too much weight to certain considerations that appear to provide reasons to change one's mind. When we are determined to be resolute and form intentions that have the special force of resolutions, we try to pre-empt the process of reconsideration that contrary inclinations might otherwise trigger by preparing ourselves, reflectively, in advance.

The formation of such resolutions can be viewed as a strategic escalation in the scope and reflectiveness of one's intentions when one anticipates the temptation to rationalize changing one's plans.

---

[16] One could revise a resolution because of contrary inclination without manifesting a failure to be resolute if one does not revise it because of contrary inclination of the sort anticipated. Suppose you plan to get a flu shot despite your intense dislike of injections, and so form a contrary inclination defeating intention to get to your appointment. If you abandon your intention at the last moment because you hate shots, that is a failure to be resolute. If you abandon your intention because you learn that they have run out of vaccine at this site, and you would have to take a long drive to another location, it might be unclear what role is being played by the fact that you hate shots. Since it is only your disinclination to receive shots that you have fortified yourself against, this may not be a failure to be resolute.

[17] Thomas Schelling describes a variety of "techniques of self-management," including the forming of resolutions in "The Intimate Contest for Self-Command," chapter 3 of *Choice and Consequence* (Cambridge: Harvard, 1984), pp. 57–82. In "Practical Unreason," Pettit and Smith characterize the merely continent agent as someone who "must make up for a lack of spontaneous virtue by becoming a successful tactician in the art of self-management" (p. 75).

When agents form these resolutions they manifest an intelligent kind of self-doubt. This suggests that the appearance of second order commitments of this sort is best thought of as a manifestation of self-distrust, not as a necessary condition of deliberative autonomy.[18]

This account implies that failures to be resolute never take us completely by surprise. A motivational failure can take us by surprise, sure enough, but a motivational failure of a sort that was not anticipated would not be a failure to be resolute, even if it produces action against one's current better judgment. Depression, distraction, ambivalence, grief, or denial might throw a wrench into the motivational works, but these motivational failures are not necessarily failures to be resolute, not even if they involve akrasia.

Young children certainly have the ability to make judgments about what seems best, and they certainly can fail to carry out their plans. Nonetheless, they cannot fail to be resolute since they lack the ability to predict their future motivational states and to make plans that explicitly take them into account by forming resolutions. This seems to me to be a satisfying result. Young children can be impulsive, and their judgments can vacillate in response to distractions and inducements as these become salient, but they lack the cognitive sophistication to possess this technique for engineering resoluteness in themselves.

Would it be enough for the anticipation of contrary inclination to be part of the context in which a resolution is formed and not part of its content? Yes, provided that the goal of defeating that kind of contrary inclination was taken to be a reason for forming the intention.[19] The mere fact that contrary inclination was anticipated when an intention is formed is not enough to show that the intention was formed in order to defeat contrary inclination. Imagine a smoker who reasonably believes that "willpower" alone will not keep him from smoking if cigarettes are available, so he plans to manage his circumstances to make sure that cigarettes are never available: he will not buy cigarettes and he will avoid the company of smokers. Because he thinks these means will be effective, he intends to stop smoking. But then a stranger at the bus stop offers him a cigarette and he accepts it and lights up. Since his plan to quit did not use the technique of forming a contrary inclination defeating intention to refuse cigarettes when cigarettes are available, he may be akratic in smoking, but he is not weak-willed. Should he have formed such a resolution? Has he underestimated the potential of willpower as part of his comprehensive plan? Perhaps not. If he is a compulsive smoker, then he may be wise not to waste his time with the forming of contrary inclination defeating intentions, because they will not work. Once he has reduced his dependence, such intentions may become more useful to him as techniques of self-management.

---

[18] The sort of higher order volitions discussed by Harry Frankfurt in "Freedom of the Will and the Concept of a Person," this JOURNAL, LXVIII, 1 (January 1971): 5–20, as a necessary condition of freedom of the will seem to be more plausibly characterized as specific remedies for anticipated motivational conflicts: expressions of self-doubt rather than indicators of the depth of an agent's commitments.

[19] A question from Pettit and a helpful suggestion from Louis de Rossett (on different occasions) helped me clarify my views here.

Our capacity to form resolutions makes possible a distinctive kind of practical failure: a failure to form a sufficiently demanding resolution when one is necessary in order to realize one's practical goals.[20] Suppose I carry some student essays around with me planning to do some essay grading at some point in the next few days. If the essays should be returned *soon*, then I have formed only a temporally indeterminate intention when a resolution that is quite specific about the time to act is needed to get the job done. To fail to form a resolution when one believes that one will not spontaneously decide to do what one believes one must do is also a failure of practical rationality. This could even constitute a form of akrasia.[21] If I see that a more forthright plan is necessary, I am deliberatively akratic when I fail to form the necessary resolution.

It cannot be the sole goal of practical rationality that we carry out our current intentions. Other goals—that we form intentions in line with our practical judgments, that we avoid forming intentions that we will not be able to carry out, that we draft plans that are as effective as possible in realizing our goals —have as much claim to centrality. When the plans or intentions we form acknowledge the need to resist temptation and to marshal our motivational resources, then skillful draftsmanship involves placing neither too much nor too little demand on our motivational capacities. This is why being perfectly resolute does not guarantee practical rationality. Easy to carry out plans followed perfectly (for example, a plan to exercise monthly) may fail to advance our goals, while the plans that would best promote our goals (for example, a plan to exercise daily) might be too demanding to be feasible. Sometimes there is no motivationally feasible plan that will adequately promote the goal. Shortcomings in practical rationality might be inevitable in such circumstances, but because practical rationality is a complex enterprise with more than one dimension of failure, it is both too harsh and too simple to say that agents who display such shortcomings are *irrational.* In the next section I will argue that our condemnation of weak-willed failures to be resolute rests on substantive as well as procedural grounds, and that neither ground for criticism amounts to a charge of outright irrationality.

---

[20] Hill classifies as a form of weakness of will the making of weak resolves that leave too much leeway for later opting out of the decision—"Weakness of Will and Character," pp. 125, 131. I would classify such weak resolves not as weakness of will but as a distinct kind of practical failure.

[21] Bratman discusses the possibility of a similar kind of higher-order akrasia concerning judgments that endorse acting on a certain kind of desire: one may fail to endorse the desires that one judges worthy of endorsement. See "A Desire of One's Own," this JOURNAL, C, 5 (May 2003): 221–42, p. 236.

V. WHICH FAILURES TO BE RESOLUTE CONSTITUTE WEAKNESS OF WILL?

The term *weakness of will* is strongly pejorative, and not every failure to be resolute deserves to be described as weakness of will.[22] Agents who fail to carry out an overly demanding plan fail to be resolute, but they seem quite reasonable in abandoning their defective plans. They are not without blemish, however, since every failure to be resolute involves a procedural defect: the failure of a technique of self-management, and this is so no matter how unreasonable the resolution or how reasonable the agent's failure to comply with it.

Some have argued that an akratic failure to be resolute does not involve any genuine defect in practical rationality if the agent ends up doing what he has most reason to do by abandoning the original intention. Robert Audi has observed that the akratic agent's attitudes involve a local failure of coherence that might be massively outweighed by the global or "holistic" increase in coherence that is effected when the intention is revised.[23] Such a view would find, it seems, no defect worth commenting on in non-akratic failures to be resolute.

To say that a failure to be resolute may represent only an improvement in rationality requires one to adopt a nonprocedural conception of rationality according to which what is rational is what the agent has most reason to do. It equates rationality with what could be called *rational advisability.* But this is to change the subject entirely. The course of action pursued by the weak-willed agent may well be the one that the agent has most reason, given all of his attitudes viewed as a whole, to pursue, and so is in that sense, the most advisable course for the agent to pursue. But the agent had advised himself not to pursue it and had formed a contrary inclination defeating intention not to pursue it that was defeated by contrary inclination of the sort anticipated. This is why it involves a failure of practical rationality.

---

[22] Thanks to Veronica Ponce and Michael Smith for persuasion on this point.

[23] See Audi, "Weakness of Will and Rational Action," p. 329; and also Arpaly, "On Acting Rationally against One's Best Judgment," p. 491.

Some will maintain that the defect in such cases is only in the drafting of the unwise or ill-considered resolution, and so is a failure of theoretical rationality. Against this, I want to maintain that the badly drafted resolution and its failure to defeat contrary inclination both constitute failures of practical rationality. This second defect may seem negligibly small in an isolated case, but its significance will be obvious when such failures are part of a pattern in which resolutions are routinely formed only to be abandoned once it is apparent that it would not be reasonable to carry them out. Suppose that a student were to look back and note with satisfaction that each time she had formed a resolution to stay up all night to write a philosophy essay and then abandoned that resolution in the wee hours, stupefied by fatigue, a greater coherence in her overall attitudes had been produced, one which better reflected her updated estimate of the difficulty of the task and the extent of her capacities to carry it out. Something important would be missing. A technique of self-management, the forming of resolutions, had failed to work repeatedly, and although the failure of the technique might be traced ultimately to a failure in draftsmanship, the two are not entirely independent. Often a pattern of failure is what constitutes the defect in draftsmanship. If the student had been the kind of person who thrived under such conditions, there would have been nothing wrong with her resolutions.

Since the defect in failures to be resolute just is the defeat of a resolution by contrary inclination of the sort anticipated; the defect remains even if it is a good thing, from some broader point of view, that the defeat occurred. A failure to be resolute is the failure of a technique of self-management: the formation of a contrary inclination defeating intention. A self that misuses a technique of self-management by forming an unreasonable resolution may well be managing herself more skillfully later on by revising the intention rather than by carrying it out. Because of this, failures to be resolute are not always cases of failed self-management. Perhaps failures to be resolute are typically cases of failed self-management, but they are necessarily only the failure of a certain technique of self-management.

To describe a failure to be resolute as a case of weakness of will seems to involve a stronger form of condemnation. I suggest that the best account of the pejorative force in the term *weakness of will* characterizes it as a failure to be resolute that is criticizable because the agent had good reason to carry out the resolution. It is impossible to provide a schematic account of the nature of these grounds for criticism since these rest on substantive claims about what the agent had good reason to do. Still, the grounds for criticism of episodes of weakness of will do not consist in or even require the possession of akratic attitudes: a failure to be resolute can count as weakness of will without involving akrasia, and akratic failures to be resolute do not always constitute weakness of will.

It is usually assumed that akrasia is sufficient to constitute weakness of will, even without the presence of a resolution and a subsequent failure to be resolute. Holton is right, I think, to insist that weakness of will must involve a failure to carry out a resolution; a judgment that simply condemns what one is doing is not enough. A newlywed admires an attractive stranger on the bus thinking *I should not be doing this, I'm married now!* This kind of akratic "voice-over" is not enough to constitute weakness of will. The possibility of weakness of will arises only if the newlywed forms a resolution to gaze elsewhere.

Suppose that the newlywed does form a contrary inclination defeating intention to gaze elsewhere, but nevertheless continues to admire the attractive stranger. The newlywed has now failed to be resolute. Is the presence of akrasia sufficient for such a failure to be resolute to constitute weakness of will? It seems to me that it is not sufficient. Suppose the newlywed was clear-eyed and candid enough to be akratic while continuing to gaze at the attractive stranger. If the newlywed does not in fact have good reason to carry out the resolution, for example, if he or she has been overly scrupulous in working out what marital fidelity actually requires, then he or she is not weak-willed. If we are unclear about such requirements or about whether the newlywed's particular failure satisfied them or not—as I suppose we and the newlywed probably are—we will be correspondingly unclear about whether this failure to be resolute constituted weakness of will.

Akrasia need not be present for a failure to be resolute to count as weakness of will. If the newlywed is the sort of person who will find marital fidelity a great challenge and has good reason to think that admiring the attractive stranger may well constitute the first step toward disaster, and forms the resolution with this in mind, then we might judge that the newlywed is indeed weak-willed in failing to be resolute. It seems to me that this is so even if the newlywed revises or abandons the original judgment while gazing at the attractive stranger and is not akratic. Akratic attitudes are not necessary for weakness of will.

It might be objected that the presence of akratic attitudes must be sufficient to make a failure to be resolute count as weakness of will: if the newlywed still judges that it would be best to look elsewhere, then this judgment provides the newlywed with a purely procedural reason to carry it out.

Here it is important to distinguish between the defect that exists in any case: the newlywed's failure to be resolute, and the newlywed's dilemma, faced with conflicting attitudes and unable to choose between them. The newlywed might be unsure whether the resolution expressed a hasty and overly conscientious response to a harmless inclination or a well-justified sensitivity to the new and unfamiliar constraints that married life will impose. The fact that the original judgment that it would be best to gaze elsewhere has not been revised as the newlywed gazes at the attractive stranger shows that the newlywed is beset with conflicted and incoherent attitudes, but it does not favor one interpretation of the significance of the newlywed's failure to be resolute over the other. To declare that the newlywed has been weak-willed is to choose sides and judge that the problem lies not with the original resolution but with the failure to comply with it.

The presence of akratic attitudes may count as indirect evidence concerning what the agent has good reason to do, but the fact that an agent formed a resolution to do something does not settle that the agent had good reason to do it. Although we are inclined to give some weight to the agent's own convictions in deciding whether a failure to be resolute is criticizable, when we assess Huck Finn's resolution to turn in Jim or an anorexic's vow to eat only from the salad bar, we are also willing to override the agent's current perspective.[24] This is why the presence of akratic attitudes is not sufficient to make a failure to be resolute count as weakness of will.

VI. THE PRESUMPTION FAVORING RESOLUTENESS

I have argued that we simplify things too much if we say that agents who reasonably revise an unreasonable resolution have acted rationally. At the same time, we simplify things too much if we say that it is always irrational to fail to be resolute. There is a general presumption that it is reasonable not to reconsider or revise resolutions involving *contrary inclination defeating* intentions (CID intentions) because of contrary inclination of the sort anticipated, but this is no more than a presumption, and a defeasible one at that. We can have good reason to think that what we resolved is something that we have good reason not to do, perhaps because we have new evidence or perhaps because we have underestimated the reason-giving force of the contrary inclination itself. The presumption that generally holds in favor of resoluteness is not a rule that could easily be codified. If we tried to formulate a rule to guide us, it would not be enough to say:

---

[24] Some have argued that weak-willed actions must be distinguished from compulsive actions by considering the agent's capacity to act in accordance with her judgments; for example, Jeanette Kennett and Smith in "Philosophy and Commonsense: The Case of Weakness of Will," in Michaelis Michael and John O'Leary-Hawthorne, eds., *Philosophy in Mind* (Boston: Kluwer, 1994), pp. 141–57. It seems plausible to say that powerful inclinations that ensure the defeat of a resolution do not keep the episode from being a failure to be resolute, but might keep it from being criticizable, and as a result, might keep it from counting as weakness of will. Since resolutions are ineffective techniques for achieving self-control in the presence of powerful compulsive desires, agents who experience inner compulsion might not form such resolutions; in such cases, they might be akratic, without being eligible to be weak-willed.

> It is reasonable not to reconsider and to carry out a CID intention despite contrary inclination of the sort that was anticipated provided that no new evidence has come to hand and the contrary inclination does not point to evidence that one has reason to revise one's judgment about what it is best to do.

To capture the full pre-emptive force of forming a resolution while anticipating contrary inclination, something more like this would be needed:

> It is reasonable not to reconsider and to carry out a CID intention despite contrary inclination of the sort that was anticipated provided that no new evidence has come to hand and the contrary inclination does not point to evidence that one has reason to revise one's judgment about what it is best to do, unless the impression that new evidence has appeared or that contrary inclinations point to evidence that one has reason to revise one's judgment is plausibly attributed to the distorting effect of the contrary inclinations on one's evaluations.

Clearly the dispositions a reasonable agent must have regarding resoluteness cannot consist only of nonreflective mechanisms, nor can they involve applying a small number of rules as guides to judgment. Getting things right requires an ability to resist temptations masquerading as reasons to revise our judgments.

The home ground of weakness of will is the realm of policies adopted in the form of resolutions in order to promote self-control in the service of some end (for example, saving money, losing weight, getting more exercise, grading student essays).[25] I observed in section IV that in drafting such policies we have to balance a rule's efficacy—the good that it would promote if it were followed—with its feasibility. A perfectly feasible diet or budget might not yield much in the way of results, while a maximally efficacious diet or budget might not be a feasible one. Typically, the benefit to be gained from following a personal policy is accessible only if it is followed fairly consistently on future occasions. Therefore, a well-drafted personal policy must sustain the belief that one can comply with it in the future, since present compliance with the policy makes sense only if it is part of a larger pattern of compliance. This explains why a failure to be resolute might have some evidential bearing on the wisdom of a particular policy. In section V I maintained that a pattern of failures to comply with a policy can constitute a serious procedural defect. However, a single failure to comply with a personal policy need not be taken as evidence that it requires too much of us and should be revised. In fact, a policy designed to promote self-control may be insufficiently demanding if it is followed with unfailing obedience. George Ainslie has pointed out that a good reason to resist the drafting of a "loophole" to a policy in order to license an imminent failure to comply with it is that a tendency to grant ourselves such exemptions can reduce our expectation of future compliance. How can we assure ourselves that we will say "just this once" just once?[26] These considerations converge to support the somewhat surprising conclusion that flat-out akratic failures to comply with a policy can play an entirely benign role in the lives of rational agents by serving as an escape valve for the pressures created by policies that are worth maintaining, unamended, despite occasional failures to comply with them and despite the temptation to make exceptions to the policy to ratify giving in to the current temptation. Evaluative stability in the face of the desires that tempt one and lucidity about the fact that those desires do not constitute good reasons to revise one's policy are manifestations of practical rationality, and they can be displayed even when one fails to be resolute in following a policy.

---

[25] Such policies are similar to what George Ainslie has called personal rules: rules adopted in order to limit and restrain one's more impulsive desires for short-term rewards in order to gain some greater long-term reward. See *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person* (New York: Cambridge, 1992).

VII. IS AKRATIC WEAKNESS OF WILL SPECIALLY IRRATIONAL?

Someone who holds the standard view according to which the central defect in weakness of will rests on the irrationality of the weak-willed agent's akratic attitudes might make this observation:

> What you call a non-akratic failure to be resolute is not usually classified as weakness of will because it does not involve akrasia and it may not even be criticizable, since the presumption favoring resoluteness might be defeated in particular cases. As a result, I don't object to your claim that

---

[26] Ainslie, *Breakdown of Will* (New York: Cambridge, 2001), p. 87. For further discussion of Ainslie's stronger claim that past compliance gives us a reason to comply in the future, see Bratman, "Planning and Temptation," in Larry May, Marilyn Friedman, and Andy Clark, eds., *Mind and Morals* (Cambridge: MIT, 1996), pp. 293–310; and Alfred Mele, "Addiction and Self-control," *Behavior and Philosophy*, XXIV, 2 (Fall 1996): 99–117.

> failures to be resolute, though they involve a procedural defect, are not in themselves cases of practical irrationality. And what you call non-akratic weakness of will is not usually classified as weakness of will either, because in that case the agent revises or abandons the judgment behind his resolution, so I don't mind if that is classified as a failure to be resolute with an additional substantive defect. But let us focus on what you call akratic weakness of will, isn't that something even worse from the point of view of practical rationality than non-akratic weakness of will? And isn't it worse because it involves a third and additional defect: the irrationality of the agent's incoherent akratic attitudes?

The objector need not claim that akratic weakness of will constitutes a distinctively practical form of irrationality, as someone who holds the obedience view might claim. The objector might appeal to something quite different, what I will call the *rational connections view.* According to this, an akratic agent's attitudes are irrational simply because they manifest rational incoherence.[27] Since this feature can be manifested in the theoretical realm as well, no special distinction between practical rationality and theoretical rationality need be drawn to identify and condemn it.[28] Weakness of will would count as practical rather than theoretical irrationality only because it concerns deliberation and reasoning that support the formation and carrying out of intentions.

What is the rational connection between judging that you ought to do *x* instead of *y*, forming the intention to do *x*, and carrying it out? John Broome comments that if you judge that you ought to do *x*, and you do not intend to do *x*, then you are not entirely as you ought to be because your judgment *normatively requires* the corresponding intention.[29] Equally, if you do not believe as you think you ought to believe, you are not entirely as you ought to be. If you believe that you ought to do *x*, does it follow that you ought to form the intention to do *x* and that you are irrational if you do not? No. As Broome observes, all that follows from the bare facts stated so far is that you ought to be such that: [if you judge that you ought to do *x*, then you form the intention to do *x* and actually do *x* at the appropriate time]. This allows for the possibility that what you ought to do is to reconsider and revise your judgment that you ought to do *x*. The conclusion that you ought to form the intention to do *x* requires an additional premise: that you ought to judge that you ought to do *x*.[30]

---

[27] I say "incoherence" not "inconsistency" because, as Donald Davidson points out, the akratic agent's conditional all things considered judgment about what would be best is not *logically* inconsistent with the "all-out" unconditional judgment about what would be best that is actually carried out. He observes that "pure internal inconsistency" enters if one holds, as he supposes that we all do, "that one ought to act on what one judges best or obligatory, everything considered." Because weak-willed agents go against their own second order principles favoring continence, Davidson argues, though they have a reason for acting as they do, they have no reason for ignoring this second order principle. See "Paradoxes of Irrationality," in Paul Moser, ed., *Rationality in Action: Contemporary Approaches* (New York: Cambridge, 1990), pp. 456–57. On the view I have defended, the procedural support favoring resoluteness is a defeasible presumption, not a universal second order principle, so it is rational incoherence, and not pure internal inconsistency that is demonstrated in akratic weakness of will.

[28] Scanlon argues that agents' motivations may fail to align with their judgments about what they have reason to do and with their judgments about what they have reason to believe, so that the kind of irrationality involved can arise regarding belief as well as action—*What We Owe to Each Other*, pp. 35–40. This symmetry is questioned by Wallace in "Scanlon's Contractualism," *Ethics*, CXII (April 2002): 429–70; see pp. 438–41.

Consider the example involving Dixon from *Lucky Jim*. He believes he ought to tell his hostess that he left huge burn marks in the bedclothes and rug in his room. Dixon's belief normatively requires that Dixon form the intention to tell her. It does not follow from this belief alone that he ought to tell her; that would follow only if Dixon ought to believe that he ought to tell her. So let us add to the mix the additional judgment that Dixon ought to believe that he ought to tell his hostess that he set fire to the bedclothes. From this, according to Broome's argument, it would follow that Dixon ought to form the intention to tell her. Appealing to this kind of argument for support, the objector could add that Dixon is irrational in failing to form the intention to tell her since he believes that he ought to tell her, and this is a belief that he ought to have.

But Dixon believes that he could not get himself to tell his hostess what he has done even if he formed a resolution to do so. And this belief, it seems, normatively requires that Dixon not form the intention to tell her. An agent who forms intentions that he knows he will not carry out is not entirely as he ought to be, from the point of view of practical rationality. If this belief about his inability to tell her is a belief that Dixon ought to have because he has good evidence that it is true (let us suppose that this is plausible concerning Dixon, though it would be implausible about most of us in similar circumstances), then it follows that Dixon ought not form the intention to tell her. The objector should be willing to conclude here also that it would be irrational of Dixon to form that intention because he knows that he will not carry it out.

---

[29] Broome, "Normative Requirements," in Jonathan Dancy, ed., *Normativity* (Malden, MA: Blackwell, 2000), pp. 78–99, on p. 85.

[30] "Normative Requirements," p. 89, and "Normative Practical Reasoning," *Proceedings of the Aristotelian Society*, Supplemental Volume LXXV (2001): 175–93.

Dixon is in a predicament concerning the normative requirements of practical rationality because no matter what he does, Dixon will manifest a practical defect: either he will fail to form an intention to do what he thinks he ought to do or he will form an intention to do what he knows he will not be able to get himself to do. Either way, Dixon will not be entirely as he ought to be. Now, if we say, along with the objector, that Dixon will be irrational in either case, what additional force does this statement have?

Should Dixon place greater weight on avoiding irrationality than on avoiding other rational shortcomings? Since the only way that Dixon can escape from the dilemma is by relinquishing his judgment that he ought to tell his hostess, is he under rational pressure to do so? This, of course, is exactly what Dixon should not do; it is to his credit that he does not stop believing that he should tell his hostess. If he had recognized the incoherence in his attitudes and had responded by revising his judgment about what he ought to do to take into account what he was capable of doing at the moment, this would not be an improvement. There is, no doubt, some sense of *can* according to which *ought implies can.* But whatever sense that may be, the claim that an agent ought to do *x* does not entail the very strong claim that the agent can at this very moment form the appropriate intention and carry it out at the appropriate time. It would not justify Dixon in retracting his ought judgment. Perhaps Dixon ought to make himself into a person who is less self-absorbed so that he could, on such occasions, do what he ought to do. But as he is now, quite self-absorbed, he cannot do what he ought to do.[31] Lucid akrasia may well be Dixon's best move in the circumstances.

I take this short argument to be a *reductio* of the assumption that the rationally incoherent attitudes that Dixon displays or would display count as cases of irrationality in some very strong sense that implies that Dixon should place such great weight on avoiding them that he should make it the case, if he can, that the inconsistency is eliminated. For example, it might be claimed that an agent's attitudes are irrational in this very strong sense if they are defective in such a way that the agent's rational capacities alone should be sufficient to remedy the defect. Someone who has committed a straightforward logical fallacy might be considered irrational in this strong sense. But this is not true of Dixon. All things considered, Dixon is more rational to maintain these attitudes, despite their incoherence, than to revise them in order to eliminate the incoherence.[32] Reason alone might be sufficient to identify the defect here, but reason alone cannot provide the remedy for it.

---

[31] This is a case in which the agent's incapacity to comply with a rational requirement does not provide any reason to think that the requirement does not apply, but it does provide a reason to think that the failure to comply should not count as irrationality. See Wallace, "Three Conceptions of Rational Agency," p. 231, for a sustained discussion of possible grounds for maintaining the existence of rational requirements even in the face of such apparent "rational incapacities" as weakness of will.

Dixon's attitudes may be rationally defective and therefore irrational in some weaker sense: it might be said that Dixon is irrational *pro tanto*, irrational as such, in having these attitudes.[33] This weaker claim would support the judgment that Dixon has reason to avoid having such incoherent attitudes, but this reason to avoid the incoherence can be overridden by other considerations, for example, that the alternatives open to him would involve at least as much or even greater failures of rationality.

I will assume that the objector here has such a claim in mind: that akratic weakness of will is irrational *pro tanto*, in a way that non-akratic weakness of will is not, and that this additional irrationality makes akratic weakness of will worse from the point of view of practical rationality than non-akratic weakness of will. Let us focus then, on the set of rationally incoherent attitudes present simultaneously in an akratic failure to be resolute. It consists of the practical judgment that underwrote the agent's original resolution, and the intention that conflicts with that resolution on which the agent acts.

I have maintained that weak-willed agents who are not akratic deserve criticism because they have failed to be resolute in circumstances in which (i) the presumption favoring resoluteness holds, and

---

[32] In a similar spirit, Gilbert Harman argues in "Rationality" (chapter 1 of *Reasoning, Meaning and Mind* (New York: Oxford, 1999), pp. 9–45) that "even if you notice that A, B, C are inconsistent, there may still be reasons to accept each and it may be quite unclear which should be given up. You may not have the time or the ability to work out which should be given up or you may have more urgent matters to attend to before trying to figure out which to give up of A, B, C. In the meantime, it may very well be rational for you to continue to believe all three" (p. 19). More general arguments for the claim that rules of implication do not constitute rules of reasoning are found in chapters 1 and 2 of Harman's *Change In View* (Cambridge: MIT, 1986).

[33] This is all that is required by Scanlon's claim in "Reasons, Responsibility, and Reliance: Replies to Wallace, Dworkin, and Deigh," *Ethics*, CII (April 2002): 507–28, that "irrationality occurs when a person holds a judgment-sensitive attitude that is incompatible with his or her own assessment of the reasons that bear on having that attitude. This happens, for example, when a person...judges there to be compelling reason to abandon an intention but continues to have that intention" (p. 508). See also his discussion of "structural irrationality" in "Metaphysics and Morals," *Proceedings and Addresses of the American Philosophical Association*, LXXVII, 2 (November 2003): 7–22. For discussions of the difference between *pro tanto* and *prima facie* reasons, see Susan Hurley, *Natural Reasons: Personality and Polity* (New York: Oxford, 1989), pp. 130–34; and Shelly Kagan, *The Limits of Morality* (New York: Oxford, 1989), pp. 17–18.

(ii) they have good reason to carry out their resolutions. Although they deserve criticism for their failure to be resolute, they might not be in a position to see that this is the case because they have revised their original judgment about what would be best (or have simply abandoned it without revising it) when faced with a motivational failure. This is what distinguishes non-akratic weak-willed agents from akratic weak-willed agents, and though it explains why the non-akratic weak-willed agents are not in a position to indict themselves of irrationality, it does not show that they are doing better in some regard concerning practical rationality. And there is a perspective from which they seem to be doing worse. Often we are not in a good position to know whether our misgivings about carrying out some plan give us reason to reconsider or revise it. It would be better to have a tendency to be resolute in such circumstances than to have a tendency to reconsider or revise our plan. Similarly, when we do fail to carry out a resolution, it would be better to have a tendency to hold on to our original evaluative judgment in the face of uncertainty about the evidential value of our reluctance to act than to have a tendency to revise our judgment. And if we do in fact have such general tendencies that concern our evaluative judgments, it might be useful to treat uncertainty about the reasonableness of revising the judgment underlying a resolution as a reason against revising the judgment, especially when one is powerfully inclined not to carry it out. But then an akratic failure to be resolute would not be clearly worse than a non-akratic failure to be resolute in the face of such uncertainty.

The objector might concede this point, while maintaining that it holds only when we are uncertain about the credentials of the resolution that was not carried out. The objector might then press on, urging us to consider cases in which we are in an excellent epistemic position to see that the judgment underlying the resolution is one that should not be doubted and that the presumption in favor of resoluteness holds. Suppose that you have resolved to resist an inclination that you expect to be quite compelling despite the fact that there is so little to be said for it that it can be dismissed as a tempting illusion. Despite your resolution you are taken in by the strangely compelling inclination. Suppose that you were to observe this happening in yourself with perfect lucidity. Not only do you act against your better judgment here, you do so despite the fact that you are in an excellent position to see that what you judged best really is best, and that what you resolved to resist and are now failing to resist really is an illusory sort of temptation. You have failed to be resolute despite the fact that you are in an excellent position to appreciate that you have every reason to be resolute.

It is correct to say that you indict yourself of extreme, outright practical irrationality in such a case. But it is not correct to say that it is your akratic attitudes—your judgment and the perversely contrary intention that goes against it—that do the indicting here. The indicting is done by the additional beliefs and evaluative judgments that once led you to make your resolution and now undermine the credentials of the intention opposed to it. It is because you grasp the utter lack of warrant for going against your judgment that you are irrational, not simply because you act against what you currently judge to be best. This is a dramatic motivational failure that counts as an extreme form of practical irrationality, but it does not show that akratic weakness of will in such circumstances is irrational *pro tanto*, in a way that non-akratic weakness of will is not. If you had failed to be resolute non-akratically, by rationalizing and revising your practical judgment, this would have produced coherence between the new practical judgment and the intention on which you acted, but this new judgment would fail to cohere with all of the beliefs and judgments that provide such clear support for the old practical judgment. Rational incoherence amounting to outright irrationality would still occur, but in a different place. As you judged it best to do what you see very clearly that you have every reason not to do, you would indict yourself of a different form of irrationality.[34] It is a mistake to suppose, as the objector does, that structural irrationality would be present in such a case only if the agent is akratic.

The central intuition supporting the Obedience View can find a home in a broader account of practical rationality even if the Obedience View is rejected. In the presence of a vivid appreciation of the reasons to be resolute, failing to be resolute can constitute an extreme form of practical irrationality because it involves rational incoherence. And this is so whether the incoherence is manifested in akratic attitudes or rationalizing attitudes. At the same time, in the presence of powerful contrary inclinations that bring about a failure to be resolute, resisting rationalization and remaining clearheaded about one's reasons to act can constitute a modest accomplishment. Oscar Wilde allegedly said that a gentleman is someone who is never rude unintentionally.[35] Gentlemen, Wilde suggests, are never unwittingly rude. The waggish implication is that gentlemen will find themselves with reasons to be rude. My suggestion is not that rational agents will find themselves with reasons to be weak-willed, but that especially proficient rational agents will not rationalize their failures to be resolute by revising their original judgments. They will not be weak-willed unwittingly.

---

[34] The moral superiority of self-condemnation over self-justification is thoughtfully illustrated in Louise Erdrich's story "The Painted Drum," *The New Yorker* (March 3, 2003): 74–81. The narrator, an estate appraiser of Native American ancestry takes a ceremonial painted drum from a client's house and places it in her car. "What I'd just done, or was about to do, was probably a felony and could ruin our business. The ease with which I had done it bewildered me. For a person who had not stolen so much as a candy bar all her life to walk coolly out of a client's house with such a valuable object might signal insanity. The beginning of a nervous breakdown. But I didn't feel that way. I felt quite lucid. And I wondered whether others who committed irrational and criminal acts felt this calm acceptance of a previously unknown part of themselves" (p. 80). Later that night "[t]he realization that I'd stolen the drum outright surfaced and sank in my consciousness. No matter how justified by history I felt, I promised myself that I would not evade my guilt or rationalize my conduct. Which was not the same as even considering that I might do the proper thing and return the drum to Sarah Tatro" (p. 81).

ALISON MCINTYRE

Wellesley College

---

[35] In response to Cardinal Newman's comment "It is almost a definition of a gentleman to say he is one who never inflicts pain" (in *The Idea of a University*), the additional qualification "unintentionally" has been attributed to Wilde, but also to G.K. Chesterton, Oliver Herford, and to Osbert Sitwell, Edith Sitwell's brother.

BOOK REVIEWS

Deflating Existential Consequence: A Case for Nominalism. JODY AZZOUNI.
New York: Oxford University Press, 2004. vii + 342 p. Cloth $55.00.

According to a well-known theorem of Wilhelm Ackermann, there are recursive functions that grow more rapidly than any primitive recursive function. Some philosophers reject this claim on the ground that there are no functions—and indeed, no mathematical objects—of any sort. This view is sometimes called "nominalism," and in this sense of the word, Jody Azzouni is not a nominalist. Azzouni accepts the existentially quantified claims of standard mathematics (chapter 2).

And yet his new book is billed as a defense of nominalism. What is the content of Azzouni's nominalism? In his preferred formulation it is the thesis that "mathematical objects exist in no sense at all" (109). But on reflection it is clear that this cannot be his view. As Azzouni knows, in the language that Anglophone mathematicians actually speak, an existentially quantified statement like "There is at least one perfect number" is simply *equivalent*, in every relevant sense, to the corresponding existence claim: "A perfect number exists." So anyone who accepts the existentially quantified claims of mathematics must concede that there is at least *one* sense in which mathematical objects exist, namely, the sense in which the word is used in mathematics. Azzouni's claim must therefore be that mathematical objects do not exist in some other sense of the word. I propose that we introduce an unfamiliar word to express this special sense, so as to remind ourselves that the notion requires explanation. So let us say that for Azzouni, the question is not whether functions and the like exist in the ordinary, mathematical sense. The question is rather whether mathematical entities *exist**.

What might this question mean? Azzouni's book does not contain a direct answer. Azzouni assumes a reader who is gripped by—and hence familiar with—the ontological problem as he understands it. This is not Quine's problem of saying what there is. It is the problem of saying what exists in some special, freighted sense. The reader who, like this reviewer, does not already grasp this special sense is obliged, I think regrettably, to reconstruct the problem from Azzouni's discussion of it.

Azzouni sometimes suggests that the issue might be explained in familiar terms. There is a suggestion, for example, that for a thing to

0022-362X/06/0306/312–18

exist* is for it to have properties (56), and a further suggestion that an existent* thing is a thing that might be the referent of a singular term (61). So understood, Azzouni's nominalism would entail that while Ackermannn's function is recursive, it lacks the property of being recursive, or alternatively, that while there is such a thing as Ackermannn's function, the phrase "Ackermannn's function" does not denote it. But these are odd-sounding claims, and Azzouni does not pursue them.

Azzouni's dominant mode of explanation points to certain uses of the word "exists" in connection with fictional characters. Consider (F):

(F) There are fictional mice that talk, but fictional talking mice do not exist.

(F) is not obviously inconsistent, and this suggests that there is a sense of the word "exists" and a sense of the phrase "there are" in which "There are *F*s" does not entail "*F*s exist." If pressed to explain the freighted sense of "exists," Azzouni might then say: "Focus on the sense of the word in (F). When I say that mathematical objects do not exist*, I mean that they do not exist *in that sense.*"

Fair enough. But what do we learn from this explanation? As Azzouni notes (63ff.) (F) admits of several interpretations, some of which are positively unhelpful for his purposes. Thus a naïve reader might take (F) to say: *There are stories according to which there are talking mice; but those stories are false. There are no talking mice.* But then the corresponding claim about mathematics would be: *There are stories according to which Ackermannn's function exists; but those stories are false. There is no such function.* And yet it is clear that this is not Azzouni's view.

The nod towards fictional discourse thus requires supplementation. If the above paraphrase does not capture the intended reading of "*F*s do not exist*," what does? In the end Azzouni holds that while existence* cannot be defined, we do possess a workable *criterion* of existence*. We typically judge that *F*s exist* when we take them to be "ontologically independent of any linguistic or psychological process" (112–13). We say that fictional talking mice do not exist* because we suppose that they are "just made up." And this tells us something about what it means to say that *F*s exist*, even if it does not amount to a definition.

The central claim of Azzouni's nominalism is thus what might be called the *ontological dependence thesis*: There are mathematical objects of various sorts, but these objects are ontologically dependent on us. And so we ask: What does this mean? And again, Azzouni does not answer directly. Instead we get the following informal elucidation.

> The most straightforward [examples of ontological dependence] are these: characters in dreams, imaginary beings that one thinks about, fictional characters, myths and so on. And among *these* the most straightforward examples are fictional characters designed from scratch by the author for the first time. The idea is that in deciding what's true about fictional entities, the author isn't required to square what's attributed to such entities with anything. He or she can "make it all up". Perhaps it can even be said that what's true about fictional entities is only what the author stipulates as true of them. And something all of whose properties are stipulated as true isn't something that is "mind-independent" or "ontologically independent of us" (93).

This passage hints at what might be called the Simple View of fictional objects. When an author "makes up" a character *X*, he produces an *authorizing representation*, R_X, according to which *X* is *F* and *G*, and so on. According to the Simple View, when this happens the author thereby sees to it that there is such a thing as *X*, and that *X* has at least some of the properties attributed to it by its authorizing representation.[1] The larger view might then be that an object is ontologically dependent on us iff certain core truths $A(X)$ concerning *X* obtain in virtue of the fact that according to some human representation R_X, $A(X)$.[2]

Azzouni's remarks about fiction often suggest the Simple View, and if that were his position then his claim about mathematics would be tolerably clear.[3] At one point, however, he appears to repudiate the suggestion.

> I'm willing to accept as true the claim that "Sherlock Holmes is admired by many real detectives"; I'm not willing to accept as true the claim that "Sherlock Holmes lived in London in the late 19th century". We must distinguish true claims about fictional objects from those only pretended true. What takes place in fiction (usually) is pretended true (90).

This is flatly inconsistent with the suggestion, tentatively broached above, that "what's true about fictional entities is only what the author stipulates as true of them." In the end the only truth about Sherlock Holmes that Azzouni acknowledges is one that is clearly not made true by authorial stipulation.

---

[1] For familiar reasons, fictional objects cannot have *all* of the properties attributed to them in their authorizing representations. See Terrence Parsons, *Nonexistent Objects* (New York: Yale, 1980).

[2] For a neglected discussion of this view and an application to the mathematical case, see David Fair, "Representation Relative Abstract Object Idealism" (Ph.D. diss., Princeton University, 1978).

[3] For example, it might be: for each (pure) mathematical object *O* and property *M*, if *O* is *M* then *O* is *M* in virtue of the fact that, according to standard mathematics, *O* is *M*.

One way to reconcile these passages is to attribute the following view to Azzouni. Ontologically dependent objects are associated with authorizing representations. But in general they do not derive their properties from these representations. What makes X ontologically dependent is simply this: the fact that *there is such a thing as* X obtains in virtue of the fact that *according to* R_X, *there is such a thing as* X. When I dream of a poltergeist—call it P—P then exists because there is a representation according to which it exists. But P is not thereby a *poltergeist.* It remains a mere hallucination. We may say that whereas according to the Simple View, ontologically dependent items derive their *natures* from their authorizing representations, on the present view they merely derive their *being* (such as it is) from these representations.

If *this* is Azzouni's view, then his distinctive brand of nominalism comes to this: for each mathematical object X, the fact that there is such a thing as X obtains in virtue of the fact that according to some authorizing representation R_X, there is such a thing as X.

This is an intriguing view. Given plausible assumptions, it entails that there were no mathematical objects before there were mathematicians, and hence that the invocation of mathematics in the explanation of prehistoric phenomena amounts to explaining what comes earlier in terms of what comes later. This would be a surprising consequence for any philosophy of mathematics. But it is a consequence we might embrace if there were good reasons for accepting the ontological dependence thesis.

Azzouni's case for ontological dependence is a version of the old causal/epistemological argument for nominalism[4]—with a twist. Whereas the old argument was supposed to show that the existential claims of mathematics are unjustified, Azzouni's argument is designed to establish a disjunction: *either mathematical objects are ontologically dependent or the existential claims of mathematics are unjustified.* Since Azzouni rejects the second disjunct, he takes the argument to support the ontological dependence thesis.

We may reconstruct the argument as follows. Let M be the collection of our mathematical methods. Anyone who endorses these methods must regard them as reliable. And so the question arises: Why are they reliable? A *trivial* explanation says that the methods are reliable because mathematical objects are such as to have whatever properties *M* leads us to attribute to them. A *nontrivial* explanation does not have this form.

---

[4] Paul Benacerraf, "Mathematical Truth," this JOURNAL, LXX, 19 (November 8, 1973): 661–79; Hartry Field, *Realism, Mathematics, and Modality* (New York: Blackwell, 1989), pp. 35ff.

The key premise in Azzouni's argument is then as follows.

> (1) If *F*s are ontologically independent, then our beliefs about *F*s are justified only if there exists a nontrivial explanation for the reliability of our methods for arriving at them.

Azzouni then goes on to argue (mainly on the strength of a brief survey) that

> (2) There can be no nontrivial explanation for the reliability of our mathematical methods.

And given these premises, the required disjunction follows.

Let us consider the first premise first. Given a reliable method *M*, there are in fact three possibilities: there may be a nontrivial explanation of *M*'s reliability; there may be a trivial explanation, or there may be no explanation. To illustrate the third possibility, consider the various nondemonstrative methods commonly discussed under the heading of "inference to the best explanation." At present there is no generally accepted explanation for their (limited) reliability. Suppose this condition permanent. Suppose that at the end of the day we just cannot explain why the empirical theories that strike us as simple and explanatorily powerful are often true. Would this entail that insofar as our beliefs in electrons and the like are justified, these items must be reckoned ontologically dependent? Surely not. In general, if the question is whether we are justified in accepting the deliverances of a method, what matters is not whether we are in a position to explain the reliability of that method. What matters is whether it remains reasonable, all things considered, for us to regard that method as reliable. The examples suggest that it can be reasonable for us to regard a method as a reliable source of information about mind-independent reality even when, for all we know, there is no explanation for this reliability.

If there is a truth in the vicinity of Azzouni's first premise, it is this: when the correct explanation for our reliability in some area is the trivial one, then the objects in question are ontologically dependent on us. But then it is not enough for Azzouni to show that there is no nontrivial explanation for our mathematical reliability. He must argue directly for the trivial explanation. There is, however, no direct argument of this sort in Azzouni's book.

Turn now to the challenge of providing a nontrivial explanation for our reliability in mathematics. The dialectic here should exactly parallel the dialectic initiated by Hartry Field's version of the causal/

epistemological argument (*op. cit.*),[5] so we may be brief. Note first that most of the methods that mathematicians actually employ, from the grade school algorithm for long division to the appeal to infinitesimals in the early calculus, can now be given rigorous descriptions and justifications within a standard, fully rigorous, foundational framework. The challenge therefore reduces to that of explaining the reliability of the basic methods in foundational mathematics, for example, set theory. Note second that since there is no relevant mystery about the conditional reliability of mathematical proof, the real challenge will be to explain our "methods" for accepting axioms.[6] But now note, third, that it is somewhat forced to speak of "methods" in this connection. There are only a handful of axioms (or schemes) in basic set theory. Indeed, from a certain point of view there is only one: "The cumulative hierarchy exists." This axiom was accepted as the result of a complex historical process. But it would be misleading to speak of a method here on a par with sense perception or memory. Genuine methods are employed repeatedly. That is why it makes sense to ask why they are reliable. The historical process that led to the acceptance of the set theoretic axioms is, by contrast, a one-off affair. The challenge to explain the reliability of this process can only amount to a challenge to render it unmysterious that this historical process should have led to the acceptance of a body of truths.

Now for the purposes of addressing this challenge we may assume that the axioms of set theory are necessary truths. Since the challenge purports to present an internal difficulty for the mathematical realist who denies the ontological dependence thesis, we are entitled to appeal to the standard precepts of mathematical realism in responding to it. But if this is correct, then we may respond to Azzouni's challenge in several ways.

Let me mention only one. Suppose that historical investigation reveals that it is no accident that the historical process in question led to the acceptance of these axioms rather than some others. Suppose, in other words, that it turns out to be explicable that the axioms of (say) ZFC were in fact accepted. It then follows that it is no accident that our "method" led to the acceptance of a body of truths. Any method that nonaccidentally delivers a body of necessary truths is *ipso facto* a reliable method.

---

[5] For extensive discussion, see John Burgess and Gideon Rosen, *A Subject with No Object: Strategies for Nominalist Reconstrual in Mathematics* (New York: Oxford, 1997), pp. 41–49.

[6] Field, p. 231.

(It is compatible with this suggestion that if the history of mathematics had been rather different, we might have accepted different axioms. Given the role of axioms in fixing the content of the language of mathematics, it is arguable that any consistent body of axioms that we might have accepted would have been true of its intended domain. But even if the realist allows that we might conceivably have accepted a consistent theory with no natural interpretation in the (one, true) set theoretic hierarchy, this would only show that our "method" is not infallible. It would not show that it is not reliable, and indeed, explicably so.)

This is of course only the barest sketch of a response. I mention it only to suggest that the mathematical realist has ample resources for responding to Azzouni's epistemological challenge. And if this is right, Azzouni's case for his most distinctive claim about mathematics—the ontological dependence thesis—must be reckoned unpersuasive.

GIDEON ROSEN

Princeton University

THE JOURNAL OF PHILOSOPHY

BACK ISSUE SALE

Complete volumes
are now available
at the following discounted prices
in order to reduce our inventory:

Volumes 1 through 98	$10 each
Volumes 99 through 101	$25 each

For more information, contact Pamela Ward,
Business Manager, at pcw1@columbia.edu

Visa, Mastercard, American Express and
PayPal are accepted through our website at
www.journalofphilosophy.org

UNIVERSITY OF CALIFORNIA PRESS

PHILOSOPHY REVEALED

Announcing a New Series
ANCIENT PHILOSOPHIES

Stoicism
JOHN SELLARS

A lucid, comprehensive introduction to this great philosophical school. It gives an overview of the history of the school, covers its philosophy as a system, and explores the three main branches of Stoic theory.
$45.00 cloth, $16.95 paper

Songs of Experience
Modern American and European Variations on a Universal Theme
MARTIN JAY

New in Paperback—"Few in the humanities can rival Jay's omnivorousness or match his intellectual energy."
—*London Review of Books*
$21.95 paper

At bookstores or (800) 822-6657 • www.ucpress.edu

You can now subscribe to the
Journal of Philosophy using

PayPal

To learn more about this easy new option, click on the PayPal icon at **www.jphil.org.**

NEW FROM INDIANA UNIVERSITY PRESS

New Series!

Revealing Whiteness

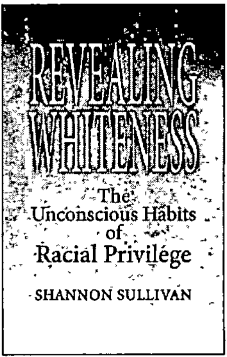
The Unconscious Habits of Racial Privilege

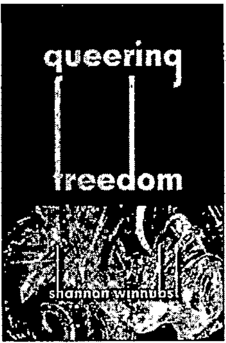
Shannon Sullivan

A bold examination of questions about whiteness and race.

American Philosophy

paper $22.95 cloth $60.00





Queering Freedom

Shannon Winnubst

Seeks individual freedom beyond rigid categories of race, gender, class, or sexual preference.

paper $21.95 cloth $55.00

The Philosophy of Simone de Beauvoir

Critical Essays

Edited by Margaret A. Simons

Reveals the depth and range of Simone de Beauvoir's thought.

A Hypatia Book

paper $24.95 cloth $65.00

Dialogue with Heidegger

Greek Philosophy

Jean Beaufret

Translated by Mark Sinclair

Heidegger discusses early Greek thinking in friendly letters to French philosopher Jean Beaufret.

Studies in Continental Thought

cloth $39.95

Indiana Series in the

Philosophy of Religion

The Weakness of God

A Theology of the Event

John D. Caputo

A challenge to theology and ontology that exposes God as weak and unstable.

paper $24.95 cloth $65.00

Heidegger's Religious Origins

Destruction and Authenticity

Benjamin D. Crowe

Sheds new light on Heidegger's early theological development.

paper $24.95 cloth $65.00

Hermeneutics at the Crossroads

Edited by Kevin J. Vanhoozer, James K. A. Smith, and Bruce Ellis Benson

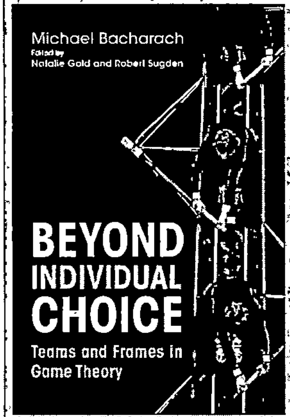
Places hermeneutics between disciplines, genres, languages, and religious commitments.

paper $24.95 cloth $65.00

800-842-6796 • iupress.indiana.edu

INDIANA

Challenging Ideas


Michael Bacharach
Edited by
Natalie Gold and Robert Sugden
BEYOND
INDIVIDUAL
CHOICE
Teams and Frames in Game Theory

Beyond Individual Choice

Teams and Frames in Game Theory

MICHAEL BACHARACH

Edited by Natalie Gold & Robert Sugden

"In this highly original book Michael Bacharach aims right at the foundations of game theory by challenging its narrow and individualistic rationality assumptions. While retaining all of the theory's analytical force, he shows how we can incorporate into it broader psychological mechanisms such as frames, social identities, and salience, producing a tool capable of biting deeper and wider into real-world social phenomena. For years to come this book will sharpen the thinking of economists. All social scientists who strive to make sense of social behavior, even those hitherto suspicious of game theory, will find it a treasure trove of inspiration."—Diego Gambetta, University of Oxford

248 pages. 7 illus. 32 tables

Cloth $35.00 0-691-12005-6

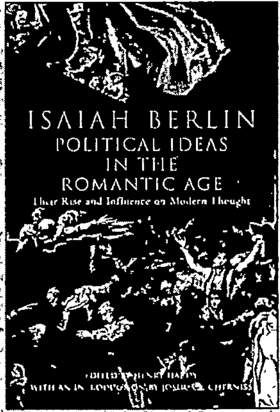
Political Ideas in the Romantic Age

Their Rise and Influence on Modern Thought

ISAIAH BERLIN

Edited by Henry Hardy

With an introduction by Joshua L. Cherniss



"*Political Ideas in the Romantic Age* makes an intriguing and provocative contribution to the history of ideas, and also to the study of Berlin's own thought. The ideas Berlin examines are intrinsically interesting and hugely influential. The book integrates Berlin's analysis of liberty with his reading of the debate between the Enlightenment and the counter-Enlightenment to an extent not found in his other works. And the editing is as meticulous as anything done by Henry Hardy, who is the best possible editor of any text by Berlin."

—George Crowder, author of *Isaiah Berlin: Liberty and Pluralism*

0-691-12687-9 Cloth $29.95 Due June

PRINCETON
University Press

800-777-4726
Read excerpts online
www.pup.princeton.edu

Philosophy | New from Chicago

Ancient & Contemporary

The Mirror of the Self

Sexuality, Self-Knowledge, and the Gaze in the Early Roman Empire

Shadi Bartsch

The Mirror of the Self

"Written with clarity and grace, *The Mirror of the Self* is a brilliant and original piece of work. Shadi Bartsch does an excellent job of bringing together literature, history, and philosophy to produce an immensely rich book. It will make a big splash in the fields of classics, comparative literature, and rhetoric."—Andrea Nightingale, Stanford University

Cloth $45.00

The Secret History of Emotion

From Aristotle's Rhetoric to Modern Brain Science

Daniel M. Gross

"With *The Secret History of Emotion*, Daniel Gross has achieved what I thought impossible: he compresses into these pages a compelling history of emotion from Aristotle to today. His argument that there exists a great tradition of understanding the emotions as a psychosocial phenomenon is cogent, coherent, and interesting from beginning to end. This is a remarkable book."
—David Konstan, Brown University

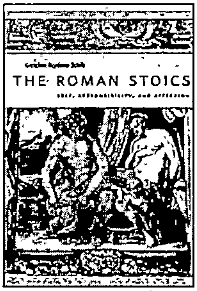
Cloth $35.00

Now in paperback

The Roman Stoics

Self, Responsibility, and Affection

Gretchen Reydams-Schils



THE ROMAN STOICS

"A clear and well-argued book. . . . A very valuable (and enjoyable) contribution to Stoic scholarship."
—*Bryn Mawr Classical Review*

Paper $26.00

Feminism and the Abyss of Freedom

Linda M. G. Zerilli

"An exciting and exhilarating book, clearly destined to have a lasting influence in feminist theory. . . . Zerilli makes a serious effort to go beyond the terrain of identity politics and its critics. The emphasis is put, on the contrary, on political judgment as the basis for forming political communities and on practices of freedom."
—Ernesto Laclau, University of Essex

Paper $22.00

New from Dana Press

Hard Science, Hard Choices

Facts, Ethics, and Policies Guiding Brain Science Today

Sandra Ackerman

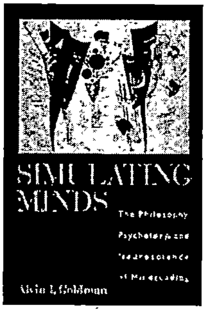
Acclaimed science writer Sandra Ackerman presents a compelling synthesis of an interdisciplinary conference considering the ethical and medical challenges posed by advances in neuroscience research.

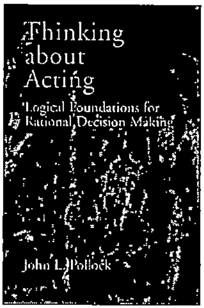
The Dana Series on Neuroethics

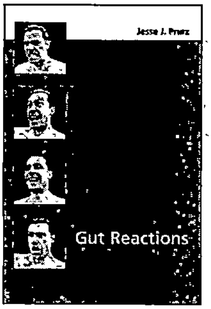
Paper $12.95

The University of Chicago Press
www.press.uchicago.edu

New from
Oxford







Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD
UNIVERSITY PRESS

Simulating Minds
The Philosophy, Psychology, and Neuroscience of Mindreading
ALVIN I. GOLDMAN

People are minded creatures; we have thoughts, feelings, and emotions. More intriguingly, we grasp our own mental states, and conduct the business of ascribing them to ourselves and others without instruction in formal psychology. How do we do this? And what are the dimensions of our grasp of the mental realm? Alvin I. Goldman explores these questions with the tools of philosophy, developmental psychology, social psychology and cognitive neuroscience.

(Philosophy of Mind Series)
June 2006 $35.00

Thinking about Acting
Logical Foundations for Rational Decision Making
JOHN L. POLLOCK

This book aims to construct a theory of rational decision making for real, resource-bounded agents. Such decision making must be based on objective probabilities rather than subjective probabilities, and cannot be done by choosing single actions with maximal expected values. Actions must be chosen as parts of plans, and plans must be evaluated in the context of other plans.

June 2006 $55.00

New in Paperback!
A Choice *Outstanding Academic Title for 2005*

Gut Reactions
A Perceptual Theory of Emotion
JESSE J. PRINZ

"Excellent...Prinz's book is utterly compelling and a valuable read for any student or researcher of the emotions, philosophy of mind, and perception."—*Metapsychology Online Book Reviews*

(Philosophy of Mind Series)
paper June 2006 $24.95

The Problem of Evil
PETER VAN INWAGEN

The vast amount of suffering in the world is often held as a particularly powerful reason to deny that God exists. Now, one of the world's most distinguished philosophers of religion presents his own position on the problem of evil. Highly accessible and sensitively argued, Peter van Inwagen's book argues that such reasoning does not hold: his conclusion is not that God exists, but that suffering cannot be shown to prove that He does not.

June 2006 $35.00

CAMBRIDGE

BEST IN SCHOLARSHIP

Nietzsche: Thus Spoke Zarathustra
Edited by Robert Pippin
Translated by Adrian Del Caro
Cambridge Texts in the History of Philosophy
$65.00: Hb: 0-521-84171-2. 316pp
$17.99: Pb: 0-521-60261-0

The Family and the Political Self
Laurence Thomas
$65.00: Hb: 0-521-85417-2: 198pp
$24.99: Pb: 0-521-67011-X

Rousseau
The Sentiment of Existence
David Gauthier
$60.00: Hb. 0-521-80976-2: 212pp
$22.99: Pb: 0-521-00759-3

Is Nature Enough?
Meaning and Truth in the Age of Science
John F. Haught
$70.00 Hb: 0-521-84714-1: 232pp
$19.99: Pb: 0-521-60993-3

The Cambridge Companion to Kant and Modern Philosophy
Paul Guyer
Cambridge Companions to Philosophy
$75.00: Hb· 0-521-82303-X: 738pp
$29.99: Pb: 0-521-52995-6

The Cambridge Companion to the Hellenistic World
Glenn Bugh
Cambridge Companions to the Ancient World
$80.00. Hb: 0-521-82879-1 402pp
$29.99. Pb. 0-521-53570-0

The Household as the Foundation of Aristotle's Polis
D. Brendan Nagle
$80.00: Hb· 0-521-84934-9: 364pp

Evolutionary Ethics and Contemporary Biology
Giovanni Boniolo and Gabriele De Anna
Cambridge Studies in Philosophy and Biology
$75 00: Hb: 0-521-85629-9: 220pp

Aristotle and the Rediscovery of Citizenship
Susan D. Collins
$70 00: Hb: 0-521-86046-6· 206pp

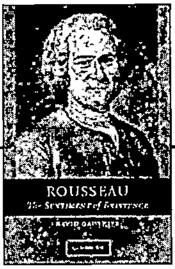
Truth, Error, and Criminal Law
An Essay in Legal Epistemology
Larry Laudan
Cambridge Studies in Philosophy and Law
$75.00: Hb: 0-521-86166-7: 254pp

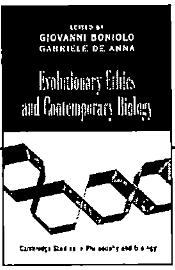
Personal Identity
Vol. 22, Part 2
Ellen Frankel Paul, Fred D. Miller, Jr, and Jeffrey Paul
Social Philosophy and Policy
$29.99: Hb: 0-521-61767-7: 400pp

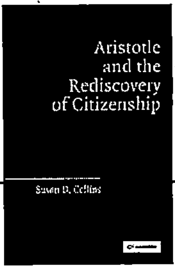
Leo Strauss and Emmanuel Levinas
Philosophy and the Politics of Revelation
Leora Batnitzky
$80.00: Hb: 0-521-86156-X: 304pp

Levinas and Theology
Michael Purcell
$70 00: Hb: 0-521-81325-5: 208pp
$27.99. Pb: 0-521-01280-5

Alcohol, Addiction and Christian Ethics
Christopher C. H. Cook
New Studies in Christian Ethics
$80.00. Hb: 0-521-85182-3: 236pp







For more information, please visit us at
www.cambridge.org/us *or call* 1-800-872-7423

CAMBRIDGE
UNIVERSITY PRESS

**prices subject to change*

New books from The MIT Press

The Parallax View

Slavoj Žižek

In Žižek's long-awaited magnum opus, he theorizes the "parallax gap" in the ontological, the scientific, and the political—and rehabilitates dialectical materialism.

448 pp. $24.95

A Voice and Nothing More

Mladen Dolar

"This remarkable study of the voice and its sonority overturns all of our preconceived notions in the most stunning fashion." — Fredric Jameson, Duke University

223 pp., 4 illus. $19.95 paper

The Evolution of Morality

Richard Joyce

"This book is a tour de force, synthesizing disparate literatures into a pleasing whole. Joyce's writing is clear, articulate, and enjoyable, and his presentation masterful."
— William D. Casebeer, Associate Professor of Philosophy, U.S. Air Force Academy

Life and Mind series • A Bradford Book
288 pp., 4 illus. $32

The Kantian Legacy in Nineteenth-Century Science

edited by Michael Friedman and Alfred Nordmann

Historians of philosophy, science, and mathematics explore the influence of Kant's philosophy on the evolution of modern scientific thought.

Dibner Institute Studies in the History of Science and Technolgy
384 pp., 17 illus. $45

To order call **800-405-1619.**

http://mitpress.mit.edu

Healing Psychiatry

Bridging the Science/Humanism Divide

David H. Brendel
foreword by T. M. Luhrmann

"This is an important contribution to the philosophy of psychiatry."
— Arnold H. Modell, M. D., Harvard Medical School

Basic Bioethics series • 208 pp. $26

Introduction to Cognition and Communication

Keith Stenning, Alex Lascarides, and Jo Calder

"Human cognition sparkles brightest in the theatre of communication. This book takes you there the 'Edinburgh Way': that is, with a sound mix of psychology, logic, linguistics, computer science, and always-lively philosophical debate."
— Johan van Benthem, University of Amsterdam and Stanford University

A Bradford Book • 616 pp., 99 illus. $45

Self-Representational Approaches to Consciousness

edited by Uriah Kriegel and Kenneth Williford

Leading theorists examine the self-representational theory of consciousness as an alternative to the two dominant reductive theories of consciousness, the representational theory of consciousness and the higher-order monitoring theory.

A Bradford Book • 592 pp., 9 illus. $40 paper

now in paperback

Adapting Minds

Evolutionary Psychology and the Persistent Quest for Human Nature

David J. Buller

"The most persuasive critique of evo-psych that I have encountered."
— *The Wall Street Journal*

552 pp. $18.95 paper

Reflexive Democracy

Political Equality and the Welfare State

Kevin Olson

"An important contribution not only to debates about welfare and citizenship, but also to the theory of capability equality and deliberative democracy."
— James Bohman, Saint Louis University

Studies in Contemporary German Social Thought series • 288 pp. $35

Re-Presenting the Good Society

Maeve Cooke

"Offers a clear and cogent articulation of a central problem in critical theory and a concrete, detailed, and imaginative proposal for how to reconcile it."
— Amy R. Allen, Dartmouth College

Studies in Contemporary German Social Thought series • 304 pp. $35

Visual Versions

Robert Schwartz

Essays on topics in the theory of vision including Berkeley's influential writings on vision, debates over competing models of perception, picture perception, and the metaphysical quagmires of the "real."

A Bradford Book • 256 pp., 10 illus. $32 paper

now in paperback

Belief's Own Ethics

Jonathan E. Adler

"A beautiful book that is exceptionally learned and rich." — Igor Douven, Ars Disputandi

A Bradford Book • 403 pp. $22.50 paper

now in paperback

Action in Perception

Alva Noë

"Bold and lucid, this book brings out the best in the philosophy of mind."
— Daniel Dennett, author of *Brainchildren, Consciousness Explained,* and *Freedom Evolves*

Representations and Mind series
392 pp., 10 illus. $19 paper

Subjectivity and Selfhood

Investigating the First-Person Perspective

Dan Zahavi

"Zahavi's book is a valuable contribution to the current interdisciplinary discussion of consciousness. In simple and direct language, he gives us a full phenomenological investigation of subjectivity and selfhood." — David Carr, Emory University

A Bradford Book • 280 pp. $36

The Philosophy of Mind

Classical Problems/Contemporary Issues

Second Edition

edited by Brian Beakley and Peter Ludlow

The only sourcebook of primary readings in philosophy of mind offering in-depth coverage of both historical works and contemporary controversies, from Aristotle, Descartes, and Ryle to Noam Chomsky, Frank Jackson, and Saul Kripke.

A Bradford Book • 1,080 pp., 45 illus. $50 paper

Good and Real

Demystifying Paradoxes from Physics to Ethics

Gary L. Drescher

"In an extraordinary tour de force, Drescher has written a powerful defense of rationalism and of a deterministic universe." — Uri Wilensky, Northwestern University, author of the *NetLogo* multiagent modeling environment

A Bradford Book • 360 pp., 37 illus. $35

now in paperback

What Is Thought?

Eric B. Baum

"A book that is admirable as much for its candor as for its ambition . . . . If What is Thought can inspire a new generation of computer scientists to inquire anew about the nature of thought, it will be a valuable contribution indeed."
— Gary Marcus, *Science*

A Bradford Book • 495 pp., 34 illus. $22 paper

COMING SOON:

THE JOURNAL OF PHILOSOPHY

ONLINE

The Journal of Philosophy will be adding
a print and online subscription option for institutions
to begin Fall 2006

SUBSCRIPTION OPTIONS

Libraries and Institutions (print and online)	$180.00
Libraries and Institutions (print only)	$100.00
Individuals	$45.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

We expect to launch this new feature with our September 2006 issue.
Please consult our website at www.journalofphilosophy.org
for more details.